IDENTIFICATION OF PUTATIVE TARGETS OF NKX2-5 IN *XENOPUS LAEVIS* USING CROSS-SPECIES ANNOTATION AND MICROARRAY GENE EXPRESSION ANALYSIS

Marcus R. Breese

Submitted to the Faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biochemistry and Molecular Biology,
Indiana University

October 2011

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Howard J. Edenberg, Ph.D., Chair

_____
Thomas D. Hurley, Ph.D.

Doctoral Committee

_____
Simon J. Rhodes, Ph.D.

June 10, 2011

_____
David G. Skalnik, Ph.D.

DEDICATION

This work is dedicated to my mom.

ACKNOWLEDGEMENTS

This work would not have been possible without the help and guidance of my thesis advisor, Howard Edenberg. He was gracious to take me on as a student when my previous advisor, Matt Grow, left the university. Even though he may have asked tough questions or wanted things done in a very specific way, he was usually right. I don't think that anyone thought that this process would take nearly as long as it did, but throughout it all, he was a great mentor, and I will forever be grateful to him.

I would also like to acknowledge my original advisor, Matt Grow, for getting me started on this crazy journey with frogs. This project took many strange turns, starting with spotted microarrays, pivoting to GeneChips, and finally ending with a lot of computational analysis. Throughout each of those steps, Matt gave me a great deal of leeway and help when I needed it. He also let me explore the bioinformatics side of science before jumping back into benchwork. Even though he left the university before the end of my work, he set me up with a solid foundation with which to continue. His enthusiasm for science was infectious, and I learned a great deal from him.

For the past year and a half, Yunlong Liu has kindly let me work in his lab while I finished this work. He let me play in the world of next-generation sequencing by day while I worked on my thesis by night (and quite often vice-versa). He has been very supportive of me, and I am quite appreciative.

I also want to thank my thesis committee members: Tom Hurley, David Skalnik, and Simon Rhodes. I am especially thankful to Dr. Rhodes for stepping in when Matt left. My

committee was always very patient with my work, allowing me the opportunity to explore the computational aspects of this research, while kindly reminding me that I was in the Department of Biochemistry and Molecular Biology and needed to finish my benchwork. Together, they helped me to get everything possible from my data.

Finally, I'd like to thank my family for putting up with me and my schedule. This work is the result of many late nights (and quite a few late mornings). My wife, Erin, has dealt with it all in stride, putting up with me in the process. Throughout the duration of this project, we got married and went from daily walks with the dog to less-frequent walks with the kids (and the dog). None of this would have been possible without her.

ABSTRACT

Marcus R. Breese

Identification of putative targets of Nkx2-5 in *Xenopus laevis* using cross-species annotation and microarray gene expression analysis

The heart is the first organ to form during development in vertebrates and Nkx2-5 is the first marker of cardiac specification. In *Xenopus laevis*, Nkx2-5 is essential for heart formation, but early targets of this homeodomain transcription factor have not been fully characterized. In order to discover potential early targets of Nkx2-5, synthetic Nkx2-5 mRNA was injected into eight-cell *Xenopus laevis* embryos and changes in gene expression measured using microarray analysis. While *Xenopus laevis* is a commonly used model organism for developmental studies, its genome remains poorly annotated. To compensate for this, a cross-species annotation database called CrossGene was constructed. CrossGene was created by exhaustively comparing UniGene transcripts from *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans* using the BLAST family of algorithms. Networks were then assembled by recursively combining reciprocal best matches into groups of orthologous genes. Gene ontology annotation from all organisms could then be applied to all members of the reciprocal group. In this way, the CrossGene database was used to augment the existing genomic annotation of *Xenopus laevis*.

Combining cross-species annotation with differential gene expression analysis of Nkx2-5 overexpression led to the discovery of 99 potential targets of Nkx2-5.

Howard J. Edenberg, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BMP	Bone morphogenic protein
BLAST	Basic Local Alignment Search Tool
Bp	Base pairs
C_0	Cycle zero, prior to amplification
C_p	Crossing point
C_t	Threshold cycle
cDNA	Complementary DNA
ChIP	Chromatin-immunoprecipitation
cRNA	Complementary RNA
DNA	Deoxyribonucleic acid
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
dGTP	Deoxyguanosine triphosphate
DMSO	dimethyl sulfoxide
dNTP	Mixture of dATP, dCTP, dGTP, dTTP
DTT	Dithiothreitol
dTTP	Deoxythymidine triphosphate
EDTA	ethylenedinitrilotetraacetic acid
EST	Expressed sequence tag
EtBr	Ethidium bromide
FDR	False discovery rate
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GFP	Green fluorescent protein
GO	Gene Ontology

GOA	Gene Ontology Annotation database from EBI
HA	Human influenza hemagglutinin epitope
HCG	Human chorionic gonadotropin
HCl	Hydrochloric acid
HBOX	Homeobox
IEA	GO annotation that was inferred using electronic annotation
IPA	Ingenuity Pathway Analysis software
LB	Luria broth
MMR	Mark's modified Ringer's solution
mRNA	Messenger RNA
NaCl	Sodium chloride
NCBI	National Center for Biotechnology Information
NKE	Nkx2-5 enhanced binding site
ODC	Ornithine decarboxylase
ORF	Open reading frame
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
SDS	Sodium docecyl-sulfate
Taq	*Thermus aquaticus*
TGF-β	Transforming growth factor beta
Tris-HCl	Tris base, pH balanced with HCl
UTR	Untranslated region
UV/Vis	Ultraviolet / visible light

CHAPTER 1: INTRODUCTION

Cardiogenesis

The heart is the first major organ to develop and it does so via a well-coordinated series of events including timed changes in gene expression and cellular migration. The mechanisms of heart development are similar for all vertebrates, indicating that the developmental mechanisms are highly conserved evolutionarily. Indeed, the mechanisms for heart development are so well conserved that much of the early work in the field was performed by studying the formation of the *Drosophila* equivalent of the vertebrate heart, the dorsal vessel (Zaffran et al. 2002). Cells first become specified to the cardiac lineage soon after gastrulation (Srivastava et al. 2000) when mesoderm cells migrate laterally to form a cardiac field (or crescent) (Harvey et al. 2002).

Subtle mutations in cardiogenic genes can have a profound effect on the formation of the heart. Some of these mutations result in congenital heart disease. Congenital heart defects are the most common cause of death for infants, amounting to almost one third of all deaths due to a congenital condition (Lloyd-Jones et al. 2010). It is estimated that heart defects are present in nearly 1% of live births, of which approximately 2.3 out of 1000 will require some form of invasive treatment (Lloyd-Jones et al. 2010). Defects can range from asymptomatic ventricular septal defects that resolve themselves spontaneously to more major anatomical defect that require surgical intervention, including tetralogy of Fallot, transposition of the great arteries, atrioventricular defects, and severe ventricular septal defects. Mutations in several genes have been directly implicated in congenital heart disease in humans, these genes include Nkx2-5 (Schott et al. 1998; Benson et al.

1999), TBX5 (Basson et al. 1999), and Jagged1 (Krantz et al. 1999). Studies in *Mus musculus*, *Xenopus laevis*, *Danio rerio*, and other organisms, have uncovered mutations in additional genes that have been linked to cardiac malformations; these include TGF-β (Brown et al.), GATA4 (Kuo et al. 1997; Molkentin et al. 1997), GATA5 (Reiter et al. 1999), dHand (Srivastava et al. 1997), Nkx2-5 (Schott et al. 1998), Smad6 (Galvin et al. 2000), and Pax3 (Conway et al. 1997a; Conway et al. 1997b). Nkx2-5 is particularly interesting as it is the most commonly mutated gene in congenital heart disease (Schott et al. 1998). Recently, it has been shown that the expression of Nkx2-5 is significantly increased in patients with hypertrophic cardiomyopathy. (Kontaraki et al. 2007).

Nkx2-5

The earliest known marker of cardiogenesis in vertebrates is Nkx2-5, also known as CSX (Tonissen et al. 1994; Harvey 1996). Nkx2-5 is a homeodomain transcription factor that starts being expressed during gastrulation and continues to be expressed throughout adulthood. In vertebrates, the expression of Nkx2-5 starts in presumptive cardiac cells and continues to be restricted to the adult heart. Nkx2-5 is a member of the NK2 family of homeodomain transcription factors. It is a DNA binding protein that acts as a dimer with itself or another family member (Kasahara et al. 2001). Nkx2-5 has two DNA binding domains: a homeodomain that binds the sequence TYAAGTG and an Nk2 domain that binds the sequence CWTAATTG (Chen et al. 1995). In some known targets, such as the gene atrial natriuretic factor (ANF), the two binding sites are in close proximity in what is known as an Nk2 enhanced element (NKE) (Small et al. 2003).

A common name for Nkx2-5 is *tinman*, due to its orthology to the *Drosophila melanogaster* gene *tinman* (Tonissen et al. 1994; Evans et al. 1995). In *Drosophila melanogaster*, *tinman* is required for the formation of the insect equivalent of the heart – the dorsal vessel (Bodmer 1993). *Tinman* is named after the character in Baum's The Wonderful Wizard of Oz, because when the gene is knocked out, the organism lacks a heart, like the Tin Man in the story (Bodmer 1993). *Drosophila tinman* directly regulates other cardiac related factors, such as myocyte enhancer factor-2 (Mef2) (Gajewski et al. 1998).

In vertebrates, the role of Nkx2-5 isn't so clear. At least ten Nkx2 family members have been identified across many vertebrate species. In *Xenopus laevis*, Nkx2-3, Nkx2-5, and Nkx2-10 are all expressed in the heart field (Sparrow et al. 2000). Overexpression of Nkx2-5 in *Xenopus laevis* causes a large-heart phenotype (Cleaver et al. 1996; Harvey 1996). However, knocking down Nkx2-5 in *Xenopus laevis* or *Danio rerio* using a gene-specific morpholino oligonucleotide causes cardia bifida, but no loss of the heart organ, in contrast to *Drosophila* (Nagao et al. 2008; Tu et al. 2009). This is thought to be due to functional redundancy between the various family members (Fu et al. 1998; Grow et al. 1998). In order to test this, a dominant negative mutant was developed: Nkx2-5LP (Grow et al. 1998) (Figure 1.1). Nkx2-5LP does not effectively bind DNA, rendering it incapable of directing cardiogenic transcription. Furthermore, because Nkx2-5 operates as a heterodimer, the functional redundancy afforded by other family members was also blocked.

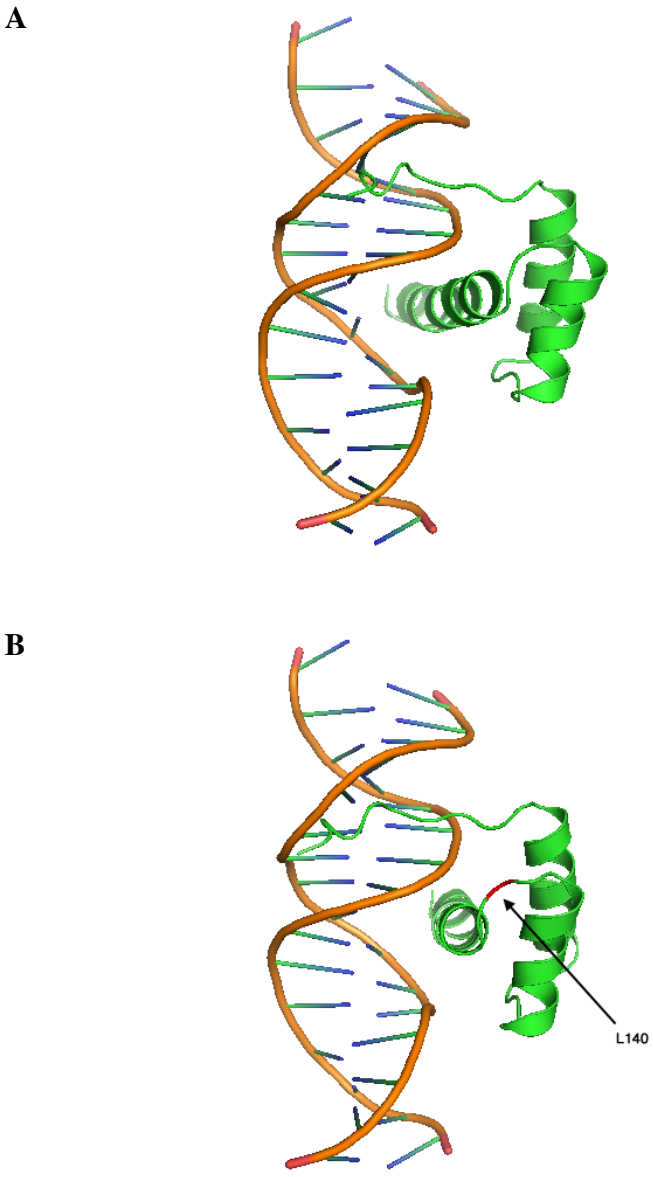


Figure 1.1 – Location of amino-acid change in the homeodomain of Nkx2-5LP dominant negative

A) *Drosophila melanogaster* vnd/NK-2 homeodomain protein bound to DNA (PDB: 1NK3) (Gruschus et al. 1997) (rendered using pymol) (DeLano Scientific 2009). B) Location of leucine to proline substitution in Nkx2-5LP is shown in red. In Nkx2-5, this substitution results in the total loss of cardiac tissue (Grow et al. 1998).

In *Mus musculus*, Nkx2-5 isn't required for cardiac specification, as it is in *Xenopus laevis*. Nkx2-5 knockouts are embryonic lethal at day 9.5-11.5 in the mouse – not because of the lack of cardiac tissue, but rather due to improper looping of the heart tube (Lyons et al. 1995). However, in murine P19 carcinoma stem cells, over-expression of Nkx2-5 is enough to drive the cells to differentiate into the cardiac lineage (Jamali et al. 2001).

Nkx2-5 is auto-regulatory, meaning that it can regulate its own expression though a positive feedback loop (Oka et al. 1997). Nkx2-5 is also known to directly interact with a number of other gene products, including GATA4 (Durocher et al. 1997; Riazi et al. 2009) and Tbx5 (Bruneau et al. 2001; Hiroi et al. 2001) to regulate the transcription of genes specific to cardiomyocytes (Figure 1.2). Examples of these targets are α-cardiac actin, ANF and myosin light chain 2 (MLC2) (Sepulveda et al. 1998; Tanaka et al. 1999). Many of these targets are expressed only in terminally differentiated, adult, cardiomyocytes. One of the known targets of Nkx2-5 in earlier development is myocardin (Myocd), which is required for cardiomyogenesis (Ueyama et al. 2003). In *Xenopus laevis,* myocardin doesn't start to be expressed until stage 24, well after the start of Nkx2-5 expression (Small et al. 2005). The lack of knowledge about early stage targets means that the role(s) of Nkx2-5 in early development have still not been fully explored.

Other cardiogenic factors

Initial cardiogenesis patterning seems to occur in response to positive and negative morphogen gradients such as the members of the bone morphogenic protein (BMP) family, Wnt, and Wnt antagonists (Harvey et al. 2002). In addition to Nkx2-5, there are many other genes that have a role in early cardiomyocyte determination (Figure 1.2).

The TGF-β signaling pathway is one such contributor. The TGF-β signaling cascade starts with BMP4 and ultimately results in activation of SMAD1 and SMAD4 (Brown et al. 2004). SMAD4 can then interact with GATA4 to regulate Nkx2-5 expression and drive cardiogenesis (Brown et al. 2004). The role of TGF-β is further supported by experiments demonstrating that a constitutively active TGF-β receptor can result in the upregulation of cardiogenic factors (Brown et al. 2004). Like BMP4, treatment with activin can also initiate cardiomyocyte differentiation via TGF-β signaling (Ariizumi et al. 2003).

The GATA family members are also important in early cardiogenic determination. GATA family members are zinc-finger transcription factors that bind to the rough consensus sequence [AT]GATA[AG] (Molkentin et al. 2000) and all are expressed in the presumptive heart field, and exhibit an overlapping expression pattern, suggesting functional redundancy (Peterkin et al. 2005). The idea of functional redundancy is reinforced by experiments involving GATA4 deficient mice where heart development continued, apparently compensated for by an increase in GATA6 expression (Pikkarainen et al. 2004). GATA6 has also been shown to activate BMP4 in adjacent endoderm, which might be required for maintenance of Nkx2-5 expression (Peterkin et al. 2003). Nkx2-5 is

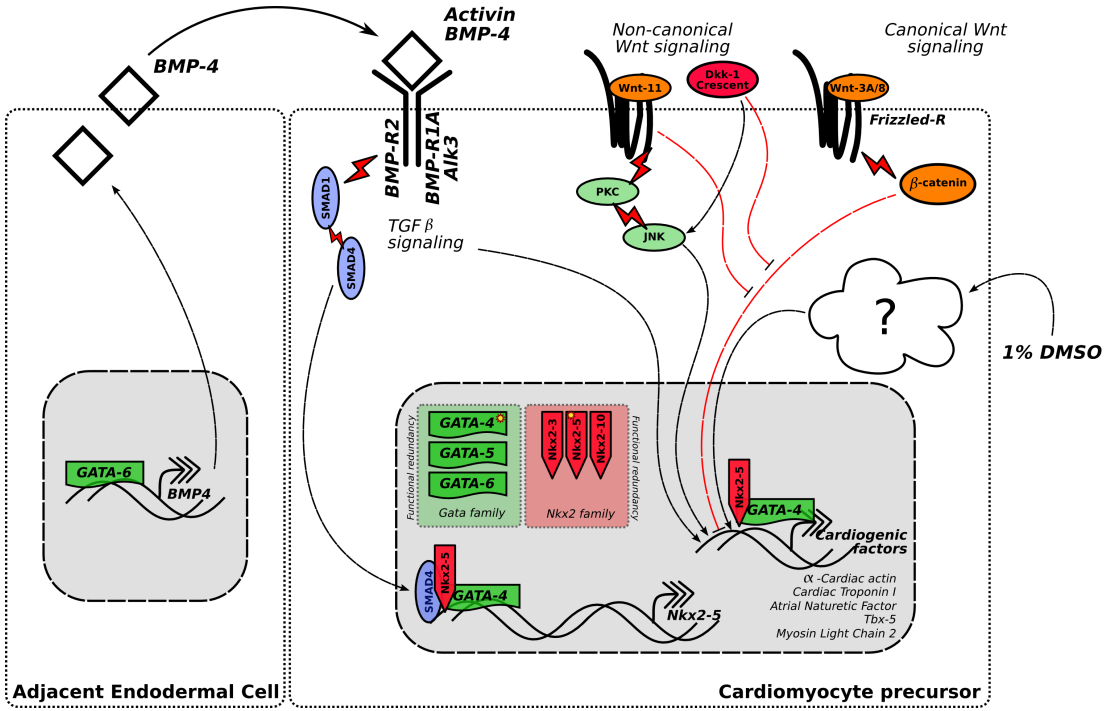


Figure 1.2 – Simplified model of known signaling in early cardiogenesis

These are some of the signaling known to be involved in early cardiogenesis. TGF-β signaling in a Cardiomyocyte precursor is activated by BMP4 signaling from adjacent endoderm. This results in SMAD4 activation which causes transcription of Nkx2-5. In addition to BMP4, exposure to activin can have the same effect. Nkx2-5, with other co-factors, regulates the transcription of terminal cardiogenic factors, such as α-cardiac actin, cardiac troponin, ANF, and others. Canonical Wnt signaling has been shown to block the transcription of cardiogenic factors through β-catenin, but non-canonical Wnt signaling has been shown to have a positive effect by either promoting the transcription of cardiogenic factors or by inhibiting the action of β-catenin. Exposure to DMSO can also cause differentiation of mouse P19 cells to cardiomyocytes, but it does so via an unknown mechanism.

known to cooperate with GATA members in activating cardiac gene expression (Durocher et al. 1997). Nkx2-5 and GATA4 have also been shown to mutually regulate each other in a positive feedback loop (Schwartz et al. 1999).

Another important pathway for cardiogenesis is Wnt signaling. Wnt signaling can be separated into two main classifications: canonical and non-canonical. Canonical Wnt signaling involves the interaction of secreted Wnt factors with the receptor Frizzled that activates β-catenin signaling. β-catenin signaling blocks cardiogenesis (Schneider et al. 2001). Additionally, treatment with the Wnt antagonists Dickkopf-1 or Crescent directly inhibits Wnt/β-catenin signaling. This inhibition then caused enhanced cardiomyocyte differentiation (Schneider et al. 2001; Pandur et al. 2002; Latinkić et al. 2003). This is in direct contrast to non-canonical Wnt signaling with Wnt-11. Wnt-11 does not activate β-catenin but instead activates PKC/JNK which also enhances cardiomyocyte differentiation (Pandur et al. 2002).

Induction of stem cells to cardiomyocytes

In addition to studying cardiogenesis in developing embryos, there are two cardiogenic stem cell induction models that should be mentioned. In these models, pluripotent stem cells are induced to form cardiomyocytes by exposure to an external factor. The most common induction model is the murine P19CL6 carcinoma cell line (McBurney et al. 1982; Habara-Ohkubo 1996). P19CL6 cells are derived from embryonal carcinoma cells and have the ability to differentiate into cardiomyocytes after exposure to 1% dimethyl sulfoxide (DMSO). After a 10-day incubation in the presence of DMSO, P19CL6 cells start to exhibit spontaneous contractions (Figure 1.2) (Habara-Ohkubo 1996). DMSO

induction requires BMP signaling via TGF β-activated kinase 1 (TAK1); however, the exact mechanism remains unknown (Monzen et al. 1999). Treatment of P19 cells with the DNA methyl transferase inhibitor 5-azacytidine has also been shown to induce cardiac differentiation (Choi et al. 2004). While the exact mechanism of 5- azacytidine is unknown, it may be an indirect effect related to altered TGF-β signaling in response to 5-azacytidine (Zuscik et al. 2004).

The other stem cell model for cardiomyocyte induction commonly used is animal cap explants from *Xenopus laevis*. Animal caps are the section of an embryo consisting of tissue from the animal pole above the blastocoel cavity. Animal caps can be extracted around stage 8-9. Because they are not exposed to the inductive signals from the vegetal region, they are left in a naïve state. If animal caps are cultured with recombinant human activin (Ariizumi et al. 2003) or injected with GATA4 mRNA (Latinkić et al. 2003), they can form spontaneously beating structures.

Use of *Xenopus laevis* in research

The African clawed frog, *Xenopus laevis*, has been a popular model organism for developmental studies for many years. The fate of each cell has been mapped and developmental staging has been well established (Faber et al. 1994). However, the biggest advantage that *Xenopus laevis* offers is the large size of their embryos. *Xenopus laevis* females can be induced to lay eggs by the injection of human chorionic gonadotropin (hCG), which was an early method for pregnancy testing in humans (Polack 1949). *Xenopus laevis* eggs are large (~1mm), which allows for easy surgical manipulation and microinjection of synthetic mRNA. Directly injecting mRNA into the

developing embryo allows a researcher to perturb a developmental gene network by over-expressing a gene, introducing a dominant-negative, or knocking out a gene using small interfering RNA (siRNA) or morpholino oligomers (Heasman et al. 2000).

This would suggest that *Xenopus laevis* would be heavily exploited in genomic studies, but this is not the case. Among 4 key model organisms that have between 10000 and 17000 PubMed references since 2000, *Xenopus laevis* has by far the fewest expression datasets in the NCBI GEO database, by factors of 4 to 18-fold (Table 1.1). One possible reason for the lack of genomic studies is that *Xenopus laevis* is an allotetraploid, having an incomplete second copy of the genome (Hughes et al. 1993; Sive et al. 2000). Thus, there are potentially four copies of a gene, two of which may be degenerate. This has made the sequencing of the *Xenopus laevis* genome and further genetic analysis difficult. However, that hasn't made the use of *Xenopus laevis* in genome-scale experiments any less desirable. Instead, other techniques were needed to compensate for the genetics of *Xenopus laevis*. Instead of relying on a fully sequenced genome to determine sequences of probes to measure gene expression, the transcriptome itself can be used. UniGene clusters from NCBI (Pontius et al. 2003; Wheeler et al. 2008) are a representation of all transcript sequences known for an organism. UniGene does not require a fully sequenced genome and is based on known mRNA sequences and unidentified EST sequences. In this way, UniGene can be used to determine probe sequences for use in microarray analysis (below).

Table 1.1 – Summary of PubMed records and GEO datasets by organism

Organism	PubMed records	% of total PubMed	GEO datasets	% of total GEO	GEO/PubMed ratio
Mus musculus	485,384	50.8%	97,993	63.6%	0.20
Rattus norvegicus	390,706	40.9%	32,046	20.8%	0.08
Gallus gallus	26,513	2.8%	2,984	1.9%	0.11
Drosophila melanogaster	17,411	1.8%	13,014	8.5%	0.75
Caenorhabditis elegans	12,233	1.3%	4,362	2.8%	0.36
Danio rerio	12,789	1.3%	2,744	1.8%	0.22
Xenopus laevis	10,522	1.1%	714	0.5%	0.07
Xenopus tropicalis	358	0.0%	110	0.1%	0.31
Total	955,916		153,967		

Date retrieved: May 4, 2011

Counts for PubMed records and GEO datasets were compiled by searching the NCBI PubMed and GEO databases. Queries were restricted to return results from only the given organism. The total numbers of records returned were taken as the record counts.

In order to compensate for the lack of functional gene annotation in *Xenopus laevis*, another approach can be used. Cross-species annotation using gene homology is one promising technique. Gene ontology (GO) terms are the standard mechanism by which the functions of genes are described (Ashburner et al. 2000b). GO terms are classified into one of three hierarchies: molecular function, biological process, and cellular component. Using these three hierarchies, it is possible to completely describe the function, role, and localization of a protein. Because *Xenopus laevis* genes are largely unannotated, the function of many genes can not readily be found. Even for well-studied genes, the annotations for *Xenopus laevis* may be lacking in many databases. For example, it is well known that Nkx2-5 is involved in heart development; however, in the UniProt Gene Ontology Annotations (GOA) database (Camon et al. 2004b), Nkx2-5 is missing the GO annotation for heart development (GO: 0007507). Indeed, the only major organism that has Nkx2-5 correctly associated to heart development using non-electronically inferred annotation is *Mus musculus*. In order to overcome this obstacle, annotations from a variety of organisms can to be assimilated to augment the existing annotation of *Xenopus laevis* genes.

The lack of a fully sequenced genome makes certain types of analysis, such as promoter analysis, impossible in *Xenopus laevis*. While the full genome sequence of *Xenopus laevis* is not available, the sequence of its close diploid cousin, *Xenopus tropicalis*, is available (Hellsten et al. 2010). The two organisms are so closely related that RNA from one organism can be readily hybridized to a cDNA library from another (Figure 1.3). Because of the close similarity between *Xenopus laevis* and *Xenopus tropicalis*, *Xenopus laevis* transcripts can be mapped to the *Xenopus tropicalis* genome. By treating the

Xenopus tropicalis genome as a surrogate for the *Xenopus laevis* genome, sequence-level analysis can be pursued.

Microarray analysis of gene expression

Microarray analysis of gene expression is the parallelization of the traditional northern blots. Northern blotting can detect the abundance of a particular RNA in a sample, using DNA or RNA probes complimentary to the target RNA molecule (Alwine et al. 1977). In northern blotting, the total RNA sample containing the target RNA is separated using electrophoresis and attached to a nitrocellulose membrane by blotting. The probes are synthesized with a detectable label, such as radioactive ^{32}P. The probes are then hybridized to the membrane and the abundance of the target RNA in the sample is then measured using autoradiography or a similar technique.

Microarray analysis is the inverse of this technique. With microarray analysis, probes targeting many genes are immobilized on a substrate in a known array pattern, allowing the detection of many genes in parallel. Once extracted from the sample of interest, the target is labeled with a detectable moiety, such as a fluorescent dye or biotin molecule. Labeled targets are hybridized to the array and measured using digital imaging (Schena et al. 1995).

Microarrays enable genome-scale gene expression experiments due to the sheer number of probes that can be present on an array. Today, a typical array can contain up to a hundred thousand probes, potentially covering all known genes for an organism. When multiple samples are compared, global changes in gene expression can be measured and

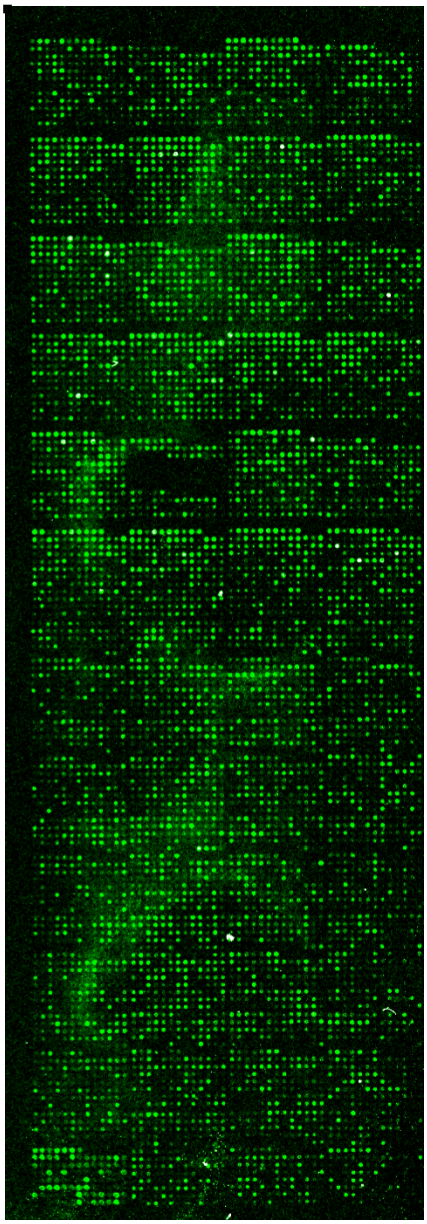


Figure 1.3 – Hybridization of *Xenopus tropicalis* heart RNA to a *Xenopus laevis* spotted cDNA microarray

This spotted microarray contains cDNA from a *Xenopus laevis* cDNA heart library and was fabricated at the Center for Medical Genomics at Indiana University School of Medicine using a VersArray ChipWriter Pro microarray robot (Bio-Rad, Hercules, CA). RNA was extracted from the pooled hearts from *Xenopus tropicalis* frogs, labeled with Cy3 dye, and hybridized to the microarray. This was then scanned using an Axon GenePix Scanner (Molecular Devices, Sunnyvale, CA). The presence of signal across a variety of probes shows that *Xenopus tropicalis* and *Xenopus laevis* have very similar gene sequences.

analyzed. One can measure gene expression in different tissues, differing stages of the cell cycle, drug dose response, or changes resulting from gene perturbation (knock-out or overexpression).

Affymetrix GeneChip arrays are a specific type of microarray in which probes are directly synthesized onto silicon wafers, using the same photolithographic techniques that are used in the semiconductor industry (Fodor et al. 1991; Lipshutz et al. 1999). In this technique, multiple probes are used in concert to detect the abundance of a single gene. By using multiple probes, closely related genes and splice variants can be measured independently. The group of multiple probes used to measure the abundance of a single target transcript is called a probe set.

One kind of GeneChip produced by Affymetrix is the *Xenopus laevis* Genome Array. This chip contains 15,503 probesets covering approximately 14,400 unique *Xenopus laevis* transcripts[1]. Transcript sequences for this GeneChip were based upon NCBI UniGene build 36 for *Xenopus laevis* (June 2003) (Pontius et al. 2003; Wheeler et al. 2008). Selection of the gene targets included on the chip was based upon annotated genes and input from the *Xenopus laevis* research community at large.

Gene Ontology

The Gene Ontology project provides a common vocabulary to describe the characteristics of genes and their functions (Ashburner et al. 2000b). The vocabulary is provided as a hierarchy of terms in parent-child relationships. These are commonly referred to as GO

[1] http://media.affymetrix.com/support/technical/datasheets/xenopus_datasheet.pdf

terms. Three different hierarchies are available: molecular function, biological process, and cellular component. The hierarchies are very flexible, allowing a term to have more than one parent term. With a controlled vocabulary, functional annotations are not limited by organism classification. With this vocabulary it is possible to extend annotations across homologous genes and proteins, even across species, in a manner that allows the accurate description of shared biology.

Individual genes can be annotated with GO terms, representing their particular molecular function, cellular location, or involvement in any biological processes. GO annotations can be derived from published experiments, manual curation, or inferred from homology to another annotated gene. Each annotation has an associated evidence code, which details the type of experimental evidence that yielded the annotation.

Scope of this work

This work has one main goal: the identification of potential targets of Nkx2-5 in early development (Chapter 2). A method for further exploring identified targets is then applied to selected targets (Chapter 3). By using *Xenopus laevis* embryos for the initial target identification, some unique data analysis challenges had to be overcome. A database was created to help overcome these challenges (Chapter 4).

Identifying targets of Nkx2-5 is important to help reveal the early signaling pathways that are critical for cardiogenesis. Chapter 2 describes one technique for studying early targets of Nkx2-5: the analysis of changes in global gene expression caused by over-expression of Nkx2-5 in whole *Xenopus laevis* embryos. Synthetic Nkx2-5 mRNA was injected into

8-cell embryos in regions from which cardiomyocytes are derived. Once the embryos reached stage 11.5, total RNA was harvested from the embryos. Stage 11.5 was chosen because it is just after endogenous Nkx2-5 starts to be expressed around stage 10, so if any other co-factors are required for binding, they should also be present. Because initial Nkx2-5 is restricted to a small subset of cells in the embryo, it was impossible to attempt to dissect out only the presumptive cardiac field. Thus, the whole embryo needed to be used for gene expression analysis. Changes in global gene expression were measured by comparing the abundance of transcripts using the Affymetrix *Xenopus laevis* Genome Array GeneChip in Nkx2-5 injected embryos and in others injected with GFP (as a control). By incorporating annotations derived from other organisms (Chapter 4), examination of gene networks and GO enrichment is possible. The end result of the microarray analysis, coupled with cross-species annotations, is a list of potential targets of Nkx2-5.

One method for implicating other genes is profiling their expression patterns among different developmental stages. The expression patterns for many genes can then be correlated to show how similar they are. This presents indirect evidence that genes may be co-regulated, but it doesn't provide direct evidence of causality. In Chapter 3, several potential targets identified in Chapter 2 are profiled using semi-quantitative RT-PCR and a subset of those were further profiled using quantitative real-time PCR. Their expression patterns were then correlated with the expression pattern of Nkx2-5.

In order to overcome some of the difficulties inherent to genomic studies in *Xenopus laevis*, a database of cross-species annotations was required. Chapter 4 describes the creation and uses of such a database. While the database itself isn't specific to *Xenopus*

laevis, it is one of the few databases of gene orthology to include *Xenopus laevis* as a supported organism. The database was constructed by finding orthologous clusters of genes from 8 eukaryotes: *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Xenopus laevis, Danio rerio, Drosophila melanogaster,* and *Caenorhabditis elegans*. Gene sequences were based on NCBI UniGene clusters for each organism (Pontius et al. 2003; Wheeler et al. 2008). These are found by clustering sequenced mRNAs and expressed sequence tags (ESTs) to form consensus gene sequences. GO annotations for each gene were then obtained from the EBI Gene Ontology Annotation (GOA) database (Camon et al. 2004b). Using the constructed network of orthologous genes, annotations from each member of the network can then be applied to the network as a whole.

CHAPTER 2: IDENTIFICATION OF PUTATIVE TARGETS OF NKX2-5 IN *XENOPUS LAEVIS*

Introduction

The heart is the first organ to form in the developing embryo. Nkx2-5 is a cardiogenic homeodomain transcription factor that is required for proper vertebrate heart development; it is the earliest known marker of the presumptive heart field (Grow et al. 1998; Schwartz et al. 1999). The binding sites of murine Nkx2-5 has been characterized by Chen and Schwartz and been shown to have two distinct binding sites: an NK2 domain and a homeodomain binding site(Chen et al. 1995). Many targets of Nkx2-5, such as ANF, myocardin, and cardiac α-actin have also been found in both late stage embryos and adults (Akazawa et al. 2005). While expression of Nkx2-5 has been well characterized at later stages, and cofactors such as GATA4 (Bruneau 2002) and Tbx5 (Durocher et al. 1998; Riazi et al. 2009) have been discovered, the role of Nkx2-5 expression very early in development remains uncertain. To find novel targets of Nkx2-5 in the early stages of development, we turned to whole embryo gene expression analysis in *Xenopus laevis*.

Over the past decade, gene expression analysis has proven to be an invaluable tool in deciphering molecular function. However, the scale of the experiments makes data analysis a rate limiting factor. A key aspect is ensuring the proper annotation of the genes. Unfortunately, the quality of annotations varies heavily from organism to organism. Even though *Xenopus laevis* is a well-studied organism, annotation of its genome is limited. As discussed in Chapter 1, a primary reason for the lack of quality

annotation is that *Xenopus laevis* is an allotetraploid, which has made traditional genetic studies using *Xenopus laevis* difficult (Sive et al. 2000). For example, in order to successfully target a gene for knock down studies, one needs to consider two (possibly degenerate) copies of the gene. We sought to compensate for the lack of annotation in *Xenopus laevis* by incorporating functional annotation from other organisms using the CrossGene database (Chapter 4). CrossGene forms clusters of similar genes using reciprocal best BLAST hits and pools GO annotations from all members of the cluster. While the genome of *Xenopus laevis* remains unsequenced, a reference assembly of the close cousin, *Xenopus tropicalis*, is available (Hellsten et al. 2010). Because of the similarity between *Xenopus laevis* and *Xenopus tropicalis, Xenopus laevis* transcripts can readily be mapped to the *Xenopus tropicalis* genome. Using the *Xenopus tropicalis* genome as a surrogate for the *Xenopus laevis* genome, sequence-level analysis is also possible. By exploiting gene annotations from other organisms and genomic sequence from *Xenopus tropicalis,* we can augment existing *Xenopus laevis* annotations to help drive data analysis.

To find potential targets of Nkx2-5 in the early stages of heart specification, *Xenopus laevis* embryos were injected with synthetic Nkx2-5 mRNA and changes in gene expression were measured using Affymetrix GeneChip microarrays. Using cross-species orthologs and annotation, we found broad changes in GO term enrichment and pathways consistent with the developmental role of Nkx2-5. Using this information and sequence-based analysis of potential Nkx2-5 binding sites, a list of likely Nkx2-5 targets was compiled. The resulting pathway analysis suggests a greater role for Nkx2-5 in the

regulation of early development and provides researchers with a list of potential Nkx2-5 targets.

Methods

Plasmid constructs

Plasmids for microinjection were previously created by cloning the coding region of the Nkx2-5 gene into the pT7Ts expression vector (Grow and Krieg, 1998). pT7Ts contains inserts from the 5' and 3' untranslated regions (UTR) from the *Xenopus laevis* β-globin gene which aids with mRNA stability and translation *in vivo*. Inserts into this plasmid can be easily transcribed into RNA *in vitro* using a T7 RNA polymerase. In microarray analysis, the lack of the Nkx2-5 3' UTR makes it possible to differentiate between endogenous Nkx2-5 mRNA and the injected mRNA. In addition to Nkx2-5, an additional construct was used, which includes an HA epitope tag at the 5' end of the Nkx2-5 gene (Nkx2-5HA) (Figure 2.1). The addition of this tag makes it possible to differentiate between endogenous Nkx2-5 RNA and injected synthetic Nkx2-5 RNA.

GFP was also previously cloned into the pCS2+ expression plasmid. This plasmid is similar to the pT7Ts plasmid, in that inserts can be *in vitro* transcribed to RNA, but in this case, SP6 RNA polymerase is used. Also, pCS2+ does not contain the extra 5' and 3' UTR inserts for *Xenopus laevis* β-globin.

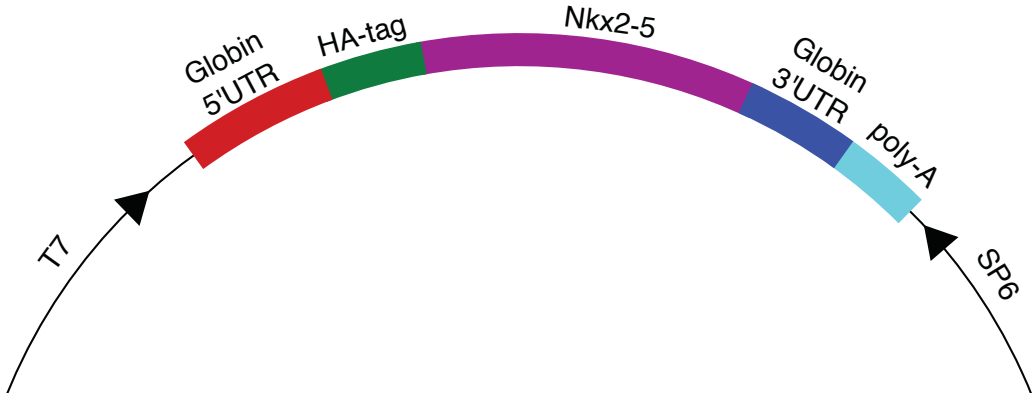


Figure 2.1 – Plasmid map of Nkx2-5HA

Nkx2-5HA is a plasmid based on the pT7Ts expression vector. It includes the coding region of Nkx2-5 as well as an in-frame HA epitope. Incorporating β-globin 5' and 3' UTRs helps to maintain stability when transcribed to RNA.

Generation of synthetic mRNA for microinjection

Synthetic mRNA was generated from the Nkx2-5HA-pT7Ts and GFP-pCS2+ plasmids using an mMessage mMachine high-yield capped RNA transcription kit (Ambion, Austin, TX), containing the appropriate RNA polymerase enzyme, T7 and SP6 respectively. Prior to mRNA synthesis, each plasmid was linearized by digesting the plasmid with the appropriate restriction enzyme. Nkx2-5HA-pT7Ts was linearized using EcoRI (Promega, Madison, WI) and GFP-pCS2+ was linearized using NotI (Promega). The digestion reaction was terminated by ethanol precipitation. RNA transcription was then performed according to the manufacturer's instructions resulting in capped, synthetic mRNAs. Synthesized mRNA was then purified using the MEGAclear purification kit (Ambion). RNA quality and yield were assessed by UV/Vis spectrophotometry using a NanoDrop spectrophotometer (Ambion) and agarose gel electrophoresis.

Culturing of Xenopus laevis embryos

Xenopus laevis embryos were collected, cultured, and manipulated using the techniques and protocols described by Sive, et al. (Sive et al. 2000). *Xenopus laevis* eggs were harvested by injecting a female frog with 400U of human chorionic gonadotropin the night before embryos were required. The following morning, eggs were harvested by inducing the female to lay eggs by physical manipulation. The eggs were collected in Petri dish and a solution of 0.1X Mark's modified Ringer's solution (MMR), pH 7.5 was added. Eggs were fertilized in vitro by mincing a small piece of freshly dissected testes in the dish. Fertilization was confirmed when the animal pole of the embryo has turned to

face the top of the dish. The protective jelly coating on the embryos was removed by rinsing the embryos in a solution of 3% cysteine until the jelly coating was sufficiently removed. Embryos were then washed multiple times in 0.1X MMR.

Embryos destined for microinjection were cultured until they reached the 4-8 cells stage (Nieuwkoop and Faber). Other embryos were allowed to develop normally in 0.1X MMR at 18-25°C until they reached desired developmental stages. These embryos were either used directly for RNA extraction or were fixed in NOTOXhisto fixing agent (Scientific Device Laboratory, Des Plaines, IL) for 2 hours (Acton et al., 2005) and stored at 4°C in methanol.

Microinjection of synthetic mRNA into Xenopus laevis embryos

Prior to injection, embryos were transferred to a 3% solution of Ficoll 400/0.1X MMR. Embryos were injected with either 250pg GFP mRNA or 250pg GFP mRNA and 250pg Nkx2-5HA mRNA. The RNA was diluted in water to yield a total injection volume of 4.6nl. GFP was used as a tracer so that it could later be determined where and if the injected RNA was actively being translated into protein. Each embryo was injected bilaterally in the dorsal vegetal blastomere with either the GFP only or GFP+Nkx2-5HA mRNA solutions using a Nanoject II (Drummond Scientific, Broomall, PA) (Figure 2.1). Embryos were cultured overnight at 15°C in the Ficoll solution. The following day, the Ficoll solution was replaced with 0.1X MMR and embryos were incubated until they reached stage 10. Then they were sorted to include only embryos that were positive for GFP expression in the cardiac crescent region (Figure 2.3).

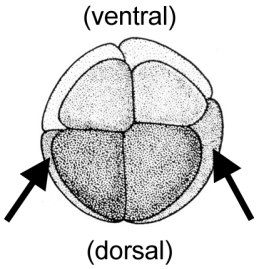


Figure 2.2 – Location of synthetic mRNA injection

An animal pole view of an 8-cell *Xenopus laevis* embryo, indicating the areas targeted for RNA injection (adapted from Faber and Nieuwkoop, 1994).

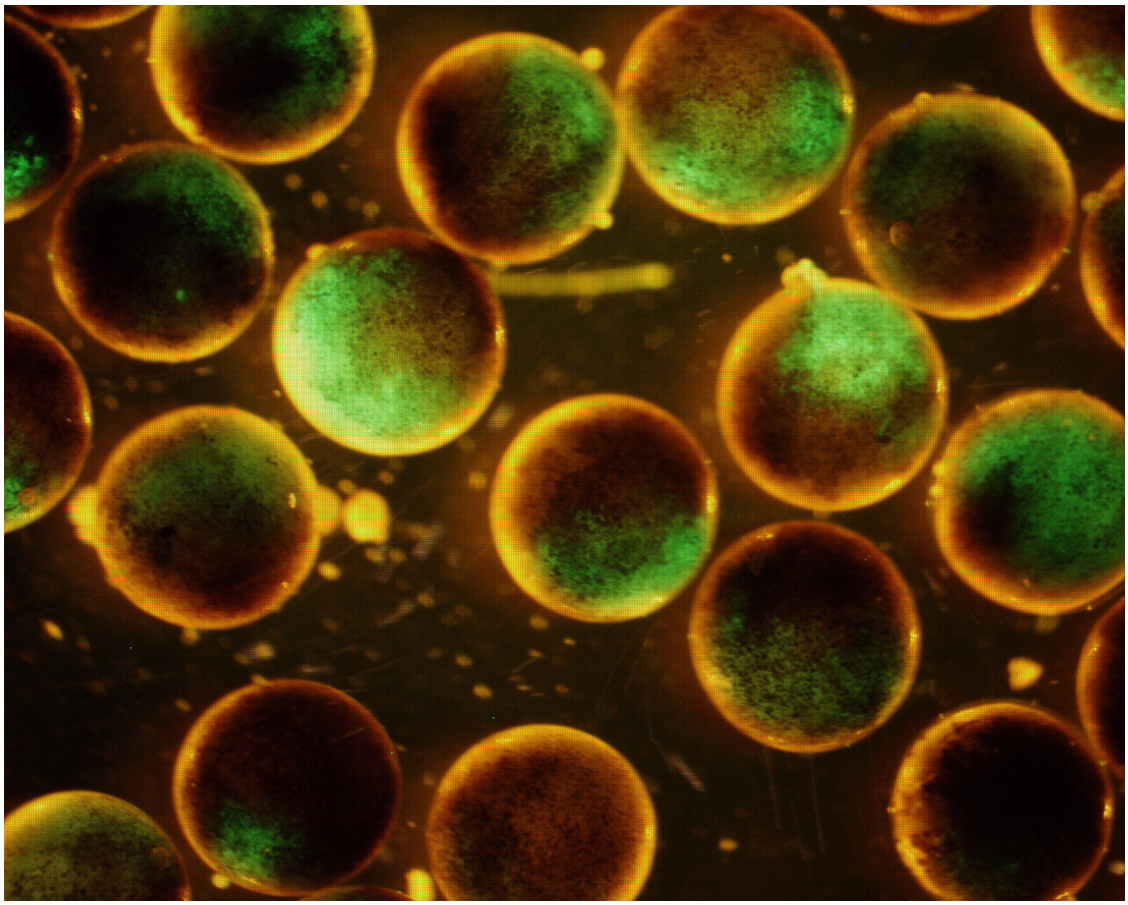


Figure 2.3 – Sorted embryos showing GFP expression in the cardiac crescent

Embryos were sorted based upon their levels of GFP expression in the cardiac crescent, indicating that injected mRNA was present in the correct region.

Harvesting RNA from Xenopus laevis embryos

RNA was extracted from individual embryos, pooled batches of stage-matched embryos, and adult tissue. Embryos that were fresh or directly stored in TRIzol reagent without further treatment were homogenized in TRIzol. Individual embryos and small pools (n<5) were homogenized using an autoclaved polypropylene microcentrifuge tube pestle (USA Scientific, Ocala, FL) prior to homogenization with an Ultra-Turrax T8 rotor/stator homogenizer (IKA Works, Wilmington, NC). Larger pools of embryos and tissues were homogenized only with the Ultra-Turrax T8 rotor/stator homogenizer. Chloroform phase extraction, followed by isopropanol precipitation was performed, following the manufacturer's instructions, resulting in isolated total RNA. For low concentration samples, 1μl of GlycoBlue glycogen (Ambion) was used as a carrier for the RNA. Isolated RNA was then cleaned up using a size appropriate RNeasy purification kit (Qiagen, Valencia, CA). RNA quality and yield were determined by UV/Vis spectrophotometry.

Reverse transcription PCR confirmation

The presence or absence of Nkx2-5HA synthetic mRNA was confirmed using non-quantitative reverse transcription PCR (RT-PCR). For each sample, 500ng of total RNA was reverse transcribed into cDNA using a SuperScript II Reverse Transcriptase kit (Invitrogen) in a 20μl reaction using anchored oligo-dT as a primer. These cDNAs were diluted to 100μl using water, resulting in solutions that contained the equivalent of 5ng/μl of the original RNA. Next, 1μl of the diluted cDNAs were amplified using a Platinum *Taq* DNA Polymerase High Fidelity PCR kit (Invitrogen) using primers specific for

ornithine decarboxylase (ODC) and Nkx2-5HA. Primers were designed algorithmically using Primer3 software with the parameters listed in table A1.1 (Rozen et al. 2000). ODC was used as an internal control. Nkx2-5HA primers were designed to be specific to that construct and to not amplify endogenous Nkx2-5 (Figure 2.4). The PCR conditions were: 94°C for 2 minutes; 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 68°C for 2 minutes; followed by 72°C for 7 minutes.

PCR products were visualized using agarose gel electrophoresis. The gel was stained with ethidium bromide (EtBr) and visualized on a 312nm UV light box. A picture was captured using a Nikon COOLPIX 995 digital camera equipped with an EtBr light filter (Nikon, Tokyo) and processed using Adobe Photoshop (Adobe, San Jose, CA).

Head versus tail dissection

Xenopus laevis embryos were cultured until they reached stage 18. Embryos were then bisected along the anterior-posterior axis into "head" and "tail" regions using an eyebrow knife (Figure 2.5)(Sive et al. 2000). RNA from three paired head and tail regions were extracted as previously described. Additionally, pools of 7 head and 7 tail samples were collected and RNA extracted. RNA expression levels from these samples were also measured using Affymetrix GeneChip microarrays.

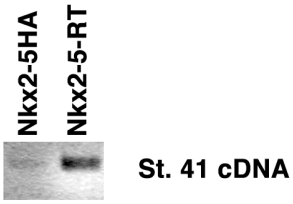


Figure 2.4 – Nkx2-5HA primers do not amplify endogenous Nkx2-5

St. 41 cDNA was used as a template for a PCR reaction with Nkx2-5HA and Nkx2-5-RT primers. After 40 cycles, using St. 41 cDNA as a template, Nkx2-5RT primers fully amplified endogenous Nkx2-5. A weak band is visible in the Nkx2-5HA lane, but this was likely a gel-loading artifact.

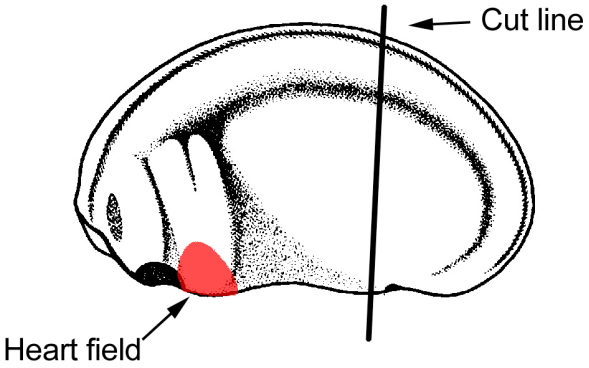


Figure 2.5 – Head versus tail bisection

The line shows the approximate location of the cut used to bisect head versus tail sections of stage 18 embryos. This is a lateral view of a stage 22 embryo and is only used for illustration (adapted from Faber and Nieuwkoop, 1994). The lower arrow points to the approximate area of Nkx2-5 expression at this stage.

Microarray analysis

Six *Xenopus laevis* embryos were injected with synthetic GFP mRNA (control samples). An additional six embryos were injected with a mix of synthetic GFP and Nkx2-5HA mRNA (experimental samples). The embryos were allowed to develop until they reached stage 11.5. At this point, total RNA was collected as previously described and cleaned up using an RNeasy micro RNA purification kit (Qiagen). Samples were treated with DNase (Qiagen) "on-column" to remove any traces of genomic DNA contamination that may have been present. RNA quality was determined using both UV/Vis spectrophotometry and a Bioanalyzer (Agilent Technologies, Santa Clara, CA). 400ng of each RNA sample were given to the Center for Medical Genomics at the Indiana University School of Medicine for processing and analysis using Affymetrix GeneChip microarrays.

Using the standard Affymetrix two-cycle amplification protocol from 2006, 100ng of each RNA sample was amplified, biotinylated, and hybridized to a *Xenopus laevis* genome array GeneChip for 17 hours at 45°C. Affymetrix Microarray Suite 5.0 was used to generate and export expression data in the form of a raw signal value, a present/absent detection call, and a p-value related to the detection call. Once exported, data were analyzed using custom scripts written in Python with NumPy (Ascher et al. 1999; Oliphant 2006) and R (Team 2010).

Using the BLASTN algorithm (Altschul et al. 1990a; Altschul et al. 1997b), the sequence of synthetic Nkx2-5HA mRNA was compared to the target sequences present on the *Xenopus laevis* genome GeneChip. No matches were found. However, the pT7Ts plasmid contains the 3'UTR sequence from *Xenopus laevis* β-globin, which caused the measured

expression level of one β-globin probeset to be artificially inflated. The synthetic GFP mRNA did not have this issue. This probeset was ignored in further analysis.

Statistical data analysis

For each probeset, mean signal values were log-transformed and a Welch's t-test performed (Welch 1947). Welch's t-test differs from Student's t-test in that it allows for unequal variances between the populations. Probesets were assigned to a *Xenopus laevis* UniGene cluster by comparing the target sequences for the probeset to the unique sequences for all *Xenopus laevis* UniGene clusters (Build 86) using the BLASTN algorithm (Altschul et al. 1990a; Altschul et al. 1997a). In all instances, the best overall match by e-value and score was used. Next, an absent/present (AP) filter was used to exclude probesets that were not marked present in 50% of either control or experimental samples (McClintick et al. 2003; McClintick et al. 2006). Probesets that did not match a *Xenopus laevis* UniGene cluster or had an e-value greater than 1e-50 were excluded. Individual probesets were considered separately even if they mapped to the same *Xenopus laevis* UniGene cluster. Next, the false discovery rate (FDR) q-value was calculated (Storey 2003). Probesets with an FDR q-value greater than 0.1, or with an absolute fold change less than 1.5 were excluded. Fold change was used as a filter to help ensure that any changes detected could be experimentally validated. Fold change was calculated by taking the ratio of the mean signal value from the Nkx2-5HA injected samples and the control GFP injected samples, with the larger of the two values used as the numerator. If the mean value of the control samples was the larger number, the value was shown as negative.

Gene ontology enrichment and annotation

Gene ontology annotations were retrieved from the CrossGene database (Chapter 4, Build 0904) for the CrossGene reciprocal best-match group corresponding to the assigned *Xenopus laevis* UniGene cluster ID. For each term, the complete hierarchy of parent terms was also retrieved and added to the list in a non-redundant manner. The complete hierarchy of terms was used for classification of genes and enrichment analysis. Enrichment or depletion of a term was calculated based upon the deviation from the expected number of genes in a group annotated with that GO term. All 13,242 probesets that had a valid *Xenopus* UniGene match with an e-value less than 1e-50 were used as the reference. Probesets that passed the 1.5 fold change filter were split into over-expressed or under-expressed groups. Terms were split into groups based upon which GO namespace they belong to: molecular function, biological process, or cellular component. This resulted in 6 individual groupings of enrichment/depletion calculations. In each group, p-values were calculated for each term using Fisher's exact test in R (Fisher 1970; Team 2010). The expected number of annotated genes was calculated based upon the percentage of genes in the reference set that were annotated with that term. Terms with an expected count of 1 were removed.

Network / pathway analysis

For each *Xenopus laevis* probeset, the *Mus. musculus* ortholog was determined by finding all *Mus musculus* genes present in the CrossGene reciprocal best-match group for the corresponding *Xenopus laevis* UniGene cluster. *Mus musculus* genes with corresponding *Xenopus laevis* fold change data were then loaded into Ingenuity Pathways Analysis

(Ingenuity Systems) and networks algorithmically calculated. In cases where multiple *Xenopus laevis* probesets mapped to the same *Mus musculus* UniGene cluster, the average of the fold changes was used. Top differentially represented biological functions were determined using a Fisher's exact test. Selected canonical pathways and interactions associated with Nkx2-5 were retrieved from Ingenuity's Knowledge Base and fold change values were overlaid.

Nkx2-5 binding site search

Murine Nkx2-5 has two binding sites: an NK2 site (TYAAGTG) and a homeobox (HBOX) site (CWTAATTG) (Transfac: M00240, M00241)(Chen et al. 1995). Using the promoter of the known *Xenopus laevis* Nkx2-5 target gene ANF, as a guide, the mouse sequences were truncated to BAAGTG and WKAAT for searching in *Xenopus* (Small et al. 2003). *Xenopus laevis* UniGene unique sequences were mapped to *Xenopus tropicalis* predicted transcripts (Assembly 4.1) using BLASTN. The single best blast hit with the lowest e-value was treated as the corresponding *Xenopus laevis* gene. Next, the 2 kb region upstream of the *Xenopus tropicalis* transcript was retrieved and searched for NK2 and HBOX potential binding sites of Nkx2-5. If two matches were within 20 bases of each other, they were treated as "paired" sites.

Results

Nkx2-5 overexpression

In order to find potential targets of Nkx2-5, 8-cell *Xenopus laevis* embryos were bilaterally injected with synthetic mRNA for Nkx2-5 with GFP mRNA as a tracer (Figure

2.6). Total RNA was extracted from whole stage 11.5 embryos that showed tracer GFP expression in the cardiac crescent region. Six Nkx2-5/GFP and six GFP-only embryos were used. RNA expression levels were measured using *Xenopus laevis* genome GeneChip microarrays (Affymetrix, Santa Clara, CA) (Figure 2.7, Figure 2.8). Results from these were then filtered based upon the Affymetrix absent/present call with a cutoff of 0.5 (McClintick et al. 2006). The false discovery rate (FDR) was then calculated and filtered using a cutoff of 0.1 (Storey 2003). Finally, we focused on probesets with changes exceeding ± 1.5 fold, resulting in a list of in 738 differentially expressed probesets covering 710 unique *Xenopus laevis* UniGene clusters (Table 2.1). Genes that were enriched in the Nkx2-5 overexpressed samples were considered up-regulated.

Development and transcription related genes enriched

To allow GO enrichment analysis of the 710 differentially expressed UniGene clusters, the *Xenopus* UniGene clusters represented on the GeneChip first needed to be annotated. Cross-species GO annotations were derived from the CrossGene database (Chapter 4) using best-match reciprocal groups. Enrichment of the 710 differentially expressed UniGene clusters was then calculated for biological process, molecular function and cellular location hierarchies using Fisher's exact test (Appendix 2) (Fisher 1970).

Analysis of up-regulated genes demonstrated that the top 10 molecular function terms were related mainly to DNA binding or transcription factor activity (Table 2.2), and in all 10 categories the up-regulated genes were over-represented, providing evidence for the role of Nkx2-5 as a positive regulator of other transcription factors. Similarly, all of the

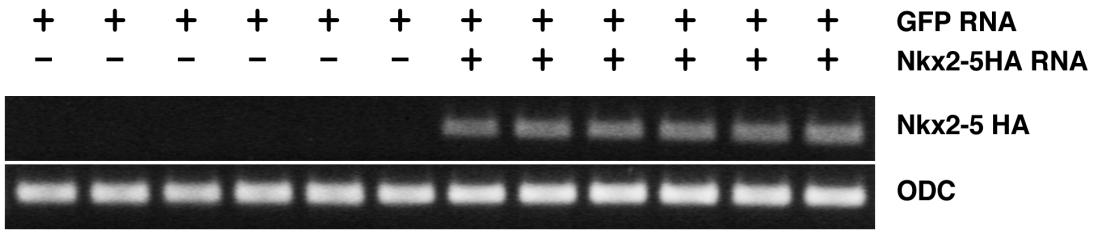


Figure 2.6 – RT-PCR confirmation of the presence of injected Nkx2-5HA RNA

Each lane represents total RNA isolated from a single embryo that was hybridized to a GeneChip. Synthetic GFP RNA was injected into all samples. Synthetic Nkx2-5HA RNA was injected into the last six samples. PCR primers specific to exogenous Nkx2-5HA were used to confirm the presence of Nkx2-5HA RNA. ODC was used as an internal control.

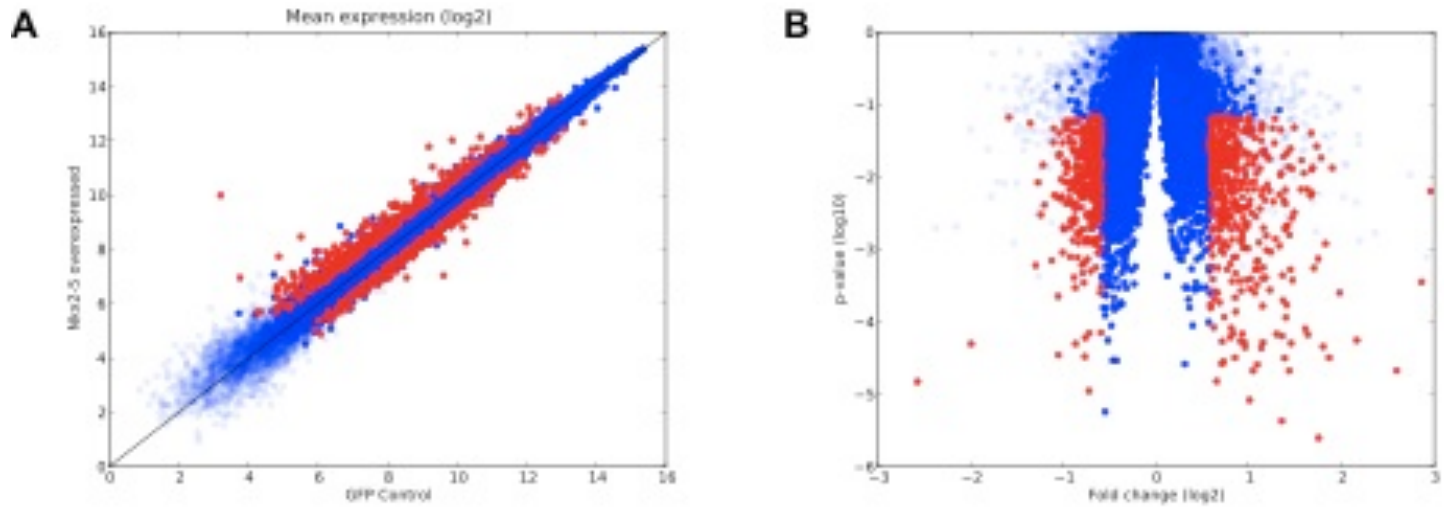


Figure 2.7 – Microarray results for Nkx2-5 over-expression samples

A) Mean expression levels (log2 transformed) for each probeset were plotted with the control (GFP) samples on the x-axis and Nkx2-5 overexpression samples on the y-axis. Probesets that passed AP, FDR and fold-change filters are in red. All others are blue with probesets that didn't pass AP filtering in light blue. B) Volcano plot showing the log2 transformed fold-change on the x-axis and the log10 transformed p-value on the y-axis. This shows the relationship between average fold change and statistical significance (p-values) for each probeset.

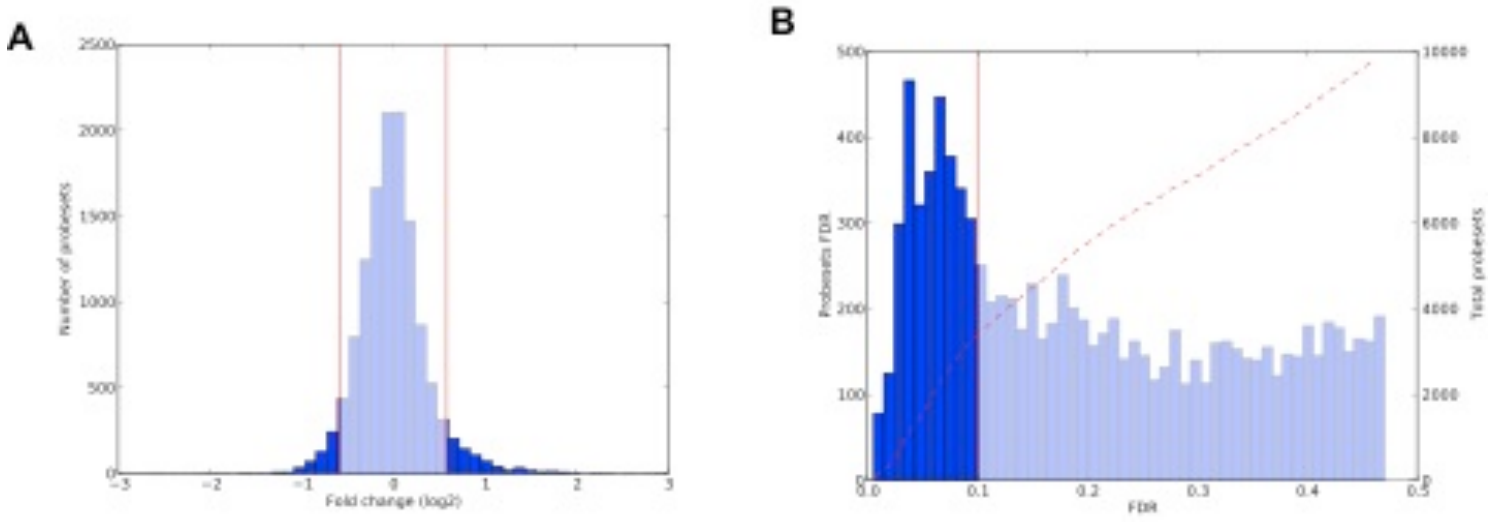


Figure 2.8 – Fold change and FDR filtering

A) Histogram showing the distribution of log2 transformed fold-change values with vertical lines showing the ± 1.5X cutoff. Light blue bins represent probesets that were excluded. B) Histogram showing the distribution of FDR q-values with the number of probesets in each bin presented on the left axis. The cumulative number of probesets is shown with the dashed red line (right axis). The FDR cutoff value of 0.1 is shown with the solid red vertical line. The intersection of the two red lines shows the number of probesets that passed the FDR filter (3,144). Light blue bins represent the probesets that were excluded.

Table 2.1 – Microarray filtering for Nkx2-5 overexpression and head vs. tail

Nkx2-5 overexpression			
Filter	***Cut off***	***Matching probesets***	***Unique UniGene***
All probesets (incl controls)		15,503	
Absent/Present Filter	(≥0.5)	11,408	
Xenopus laevis UniGene 86 match	(e-value ≤ 10^{-50})	9,818	8,845
FDR filter	(≤0.1)	3,144	2,989
Fold change (minimum)	(±1.5X)	738	709
Up-regulated	(+1.5X)	406	400
Down-regulated	(-1.5X)	332	323

Head vs. tail		
Filter		***Matching probesets***
All probesets		15,503
Absent/Present Filter	(≥0.75 in head*)	8,539
Head enriched	(>2X)	263

* In order for a probeset to be called as head-enriched, it needed to be called as "Present" in at least 3 out of 4 head samples.

top 20 biological process terms showed an excess of up-regulated genes; many of these terms relate to cell differentiation, organism patterning, organogenesis, and other developmental pathways (Table 2.3), emphasizing the role of Nkx2-5 in early development.

In contrast, analysis of down-regulated genes showed that 6 of the top 10 molecular function terms had fewer genes than expected (Table 2.4), and all 10 top biological process terms were related to development or growth and all showed fewer genes than expected (Table 2.5).

Developmental pathways activated

To examine which pathways were affected by Nkx2-5 over-expression, network analysis was performed. *Xenopus laevis* isn't supported by Ingenuity Pathways Analysis software (IPA; Ingenuity Systems, Redwood City, CA), so we mapped *Xenopus laevis* genes to *Mus musculus* orthologs using the CrossGene database, which finds the best-match reciprocal groups. For network analysis, a larger number of genes can be advantageous, so we applied an FDR cut off of 0.1 but did not filter the results by fold change, leaving 3,144 *Xenopus laevis* probesets for analysis (Table 2.1). These mapped to 2,923 unique *Mus musculus* UniGene clusters. These mouse orthologs along with the measured fold changes of the *Xenopus laevis* probesets were submitted for network analysis using IPA.

Two of the top generated networks consisted of genes with functions related to development, however, only Network 2 included Nkx2-5 itself, and it is linked to only two other genes (MEOX1 and APLNR) (Figure 2.9); NFκB is the predominant node. Network 4, while missing Nkx2-5, contains other key developmental transcription factors

Table 2.2 – Molecular function enrichment in up-regulated genes

GO Term		Expected	Actual	p-value
GO:0003700	transcription factor activity	17	33	1.6E-04
GO:0043565	sequence-specific DNA binding	13	26	5.8E-04
GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	2	7	7.1E-04
GO:0016563	transcription activator activity	11	21	2.1E-03
GO:0030528	transcription regulator activity	28	43	2.3E-03
GO:0003677	DNA binding	33	49	3.4E-03
GO:0003702	RNA polymerase II transcription factor activity	7	15	3.5E-03
GO:0046982	protein heterodimerization activity	6	13	4.3E-03
GO:0005099	Ras GTPase activator activity	2	5	6.5E-03
GO:0003723	RNA binding	16	6	9.4E-03

Table 2.3 – Biological process enrichment in up-regulated genes

GO Term		Expected	Actual	p-value
GO:0003002	regionalization	11	31	5.5E-08
GO:0007389	pattern specification process	15	37	9.7E-08
GO:0035282	segmentation	4	16	1.3E-07
GO:0030154	cell differentiation	27	53	2.6E-07
GO:0009952	anterior/posterior pattern formation	6	21	3.7E-07
GO:0032501	multicellular organismal process	55	87	4.5E-07
GO:0045893	positive regulation of transcription, DNA-dependent	13	31	8.5E-07
GO:0010557	positive regulation of macromolecule biosynthetic process	16	36	9.1E-07
GO:0051254	positive regulation of RNA metabolic process	13	31	9.4E-07
GO:0009887	organ morphogenesis	15	35	1.2E-06
GO:0001756	somitogenesis	3	12	1.6E-06
GO:0048598	embryonic morphogenesis	13	30	2.5E-06
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	10	26	2.5E-06
GO:0007275	multicellular organismal development	30	54	2.8E-06
GO:0031328	positive regulation of cellular biosynthetic process	16	36	2.9E-06
GO:0009891	positive regulation of biosynthetic process	17	36	3.1E-06
GO:0048731	system development	19	39	4.0E-06
GO:0009893	positive regulation of metabolic process	20	40	8.2E-06
GO:0048646	anatomical structure formation involved in morphogenesis	13	30	8.8E-06
GO:0010604	positive regulation of macromolecule metabolic process	19	39	9.4E-06

Table 2.4 – Molecular function enrichment in down-regulated genes

GO Term		Expected	Actual	p-value
GO:0030695	GTPase regulator activity	5	12	6.7E-04
GO:0060589	nucleoside-triphosphatase regulator activity	5	12	1.0E-03
GO:0005083	small GTPase regulator activity	3	9	2.8E-03
GO:0005198	structural molecule activity	11	2	2.9E-03
GO:0003723	RNA binding	14	5	9.5E-03
GO:0019904	protein domain specific binding	7	1	1.5E-02
GO:0008233	peptidase activity	9	2	1.8E-02
GO:0005096	GTPase activator activity	3	6	2.3E-02
GO:0004175	endopeptidase activity	6	1	3.1E-02
GO:0016887	ATPase activity	6	1	3.1E-02

Table 2.5 – Biological process enrichment in down-regulated genes

GO Term		Expected	Actual	p-value
GO:0048518	positive regulation of biological process	56	31	2.6E-05
GO:0040008	regulation of growth	33	14	9.5E-05
GO:0009653	anatomical structure morphogenesis	38	18	1.0E-04
GO:0045927	positive regulation of growth	28	11	1.3E-04
GO:0040009	regulation of growth rate	25	9	2.2E-04
GO:0040010	positive regulation of growth rate	25	9	2.2E-04
GO:0002119	nematode larval development	26	10	2.8E-04
GO:0008152	metabolic process	104	82	5.6E-04
GO:0032502	developmental process	87	65	6.6E-04
GO:0002164	larval development	26	11	6.8E-04

such as Jun, Sox2 and members of the Fox family, all of which were up-regulated by overexpression of Nkx2-5 (Figure 2.10). All of the top 5 differentially represented physiological functions were developmental (Table 2.6).

Canonical pathways and known interaction partners of Nkx2-5 were also extracted from IPA and gene expression levels overlaid as the color for each node. Two of these pathways specifically covered cardiogenesis: "Factors Promoting Cardiogenesis in Vertebrates" (Figure 2.11) and "Cardiomyocyte Differentiation via BMP Receptors" (Figure 2.12). In both of these networks, members of the BMP family were up-regulated. Out of the 69 elements known to interact with Nkx2-5, only 12 exhibited any change in expression (Figure 2.13).

Prioritization of potential Nkx2-5 targets

Classification by head/tail expression

Expression of Nkx2-5 is restricted to pre-heart tissue, so early targets should also be present in the early heart regions. However, at these early stages of development, the pre-heart fields are still migrating, making exact dissection impossible. However, the anterior half of an embryo contains the early internal organ regions, including all of the migrating pre-heart tissue. By looking for genes that are enriched in this region, it is possible to eliminate more widely expressed genes. For this, stage 18 *Xenopus laevis* embryos were bisected into "head" and "tail" sections (Figure 2.5) and RNA expression was then characterized using GeneChips (Figure 2.14). To classify a probeset as "head-enriched", the probeset had to be called present in 3 out of 4 head samples and be enriched 2-fold in

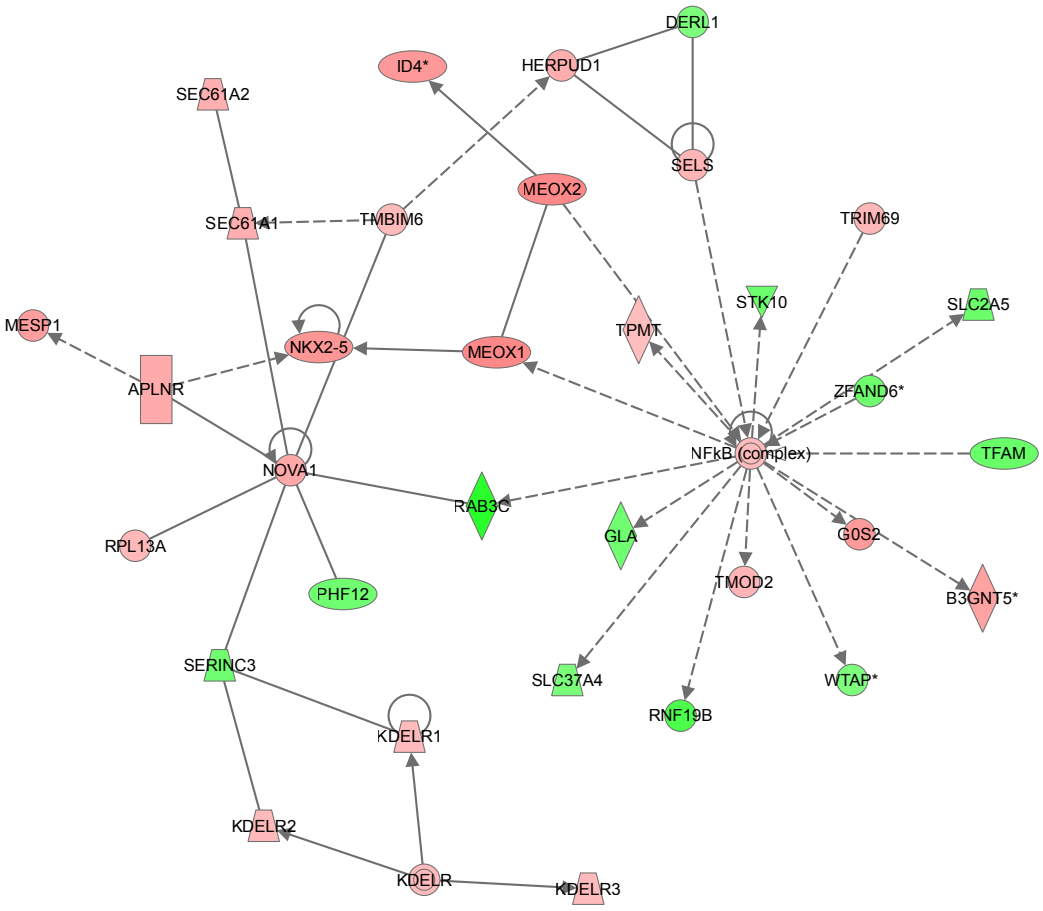


Figure 2.9 – Selected IPA Network: Embryonic Development, Tissue Development, Organismal Development

Xenopus laevis UniGene clusters were mapped to their *Mus musculus* orthologs using CrossGene best-match reciprocal groups. All genes that met a FDR cutoff of less than 0.1 were used for network generation using IPA. This is one of the networks that were generated. The color of the nodes represents the fold change of the gene, or an average where more than one *Xenopus laevis* UniGene cluster mapped to a *Mus musculus* UniGene cluster. Genes enriched in the Nkx2-5 overexpression samples are shown in red and those enriched in the GFP control samples are shown in green.

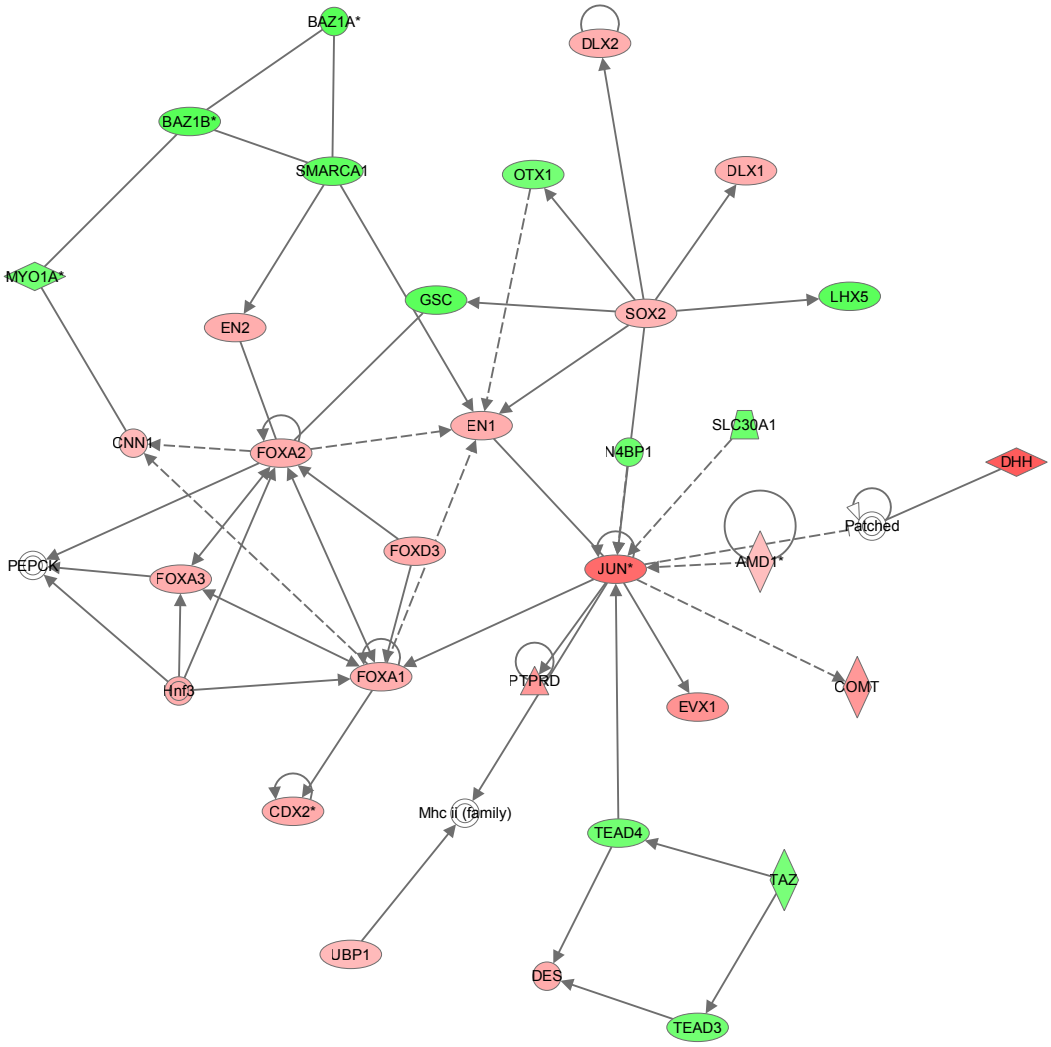


Figure 2.10 – Selected IPA Network: Cellular Development, Nervous System Development and Function, Embryonic Development

Xenopus laevis UniGene clusters were mapped to their *Mus musculus* orthologs using CrossGene best-match reciprocal groups. All genes that met a FDR cutoff of less than 0.1 were used for network generation using IPA. This is one of the networks that were generated. The color of the nodes represents the fold change of the gene, or an average where more than one *Xenopus laevis* UniGene cluster mapped to a *Mus musculus* UniGene cluster. Genes enriched in the Nkx2-5 overexpression samples are shown in red and those enriched in the GFP control samples are shown in green.

Table 2.6 – Differentially represented physiological pathways

Pathway or function	p-value range	# molecules
Embryonic Development	2.1E-15 - 2.3E-03	251
Tissue Development	2.1E-15 - 2.1E-03	211
Organismal Development	4.5E-14 – 2.0E-03	267
Cardiovascular System Development and Function	4.1E-09 – 2.0E-03	124
Organ Development	4.1E-09 - 1.2E-03	184

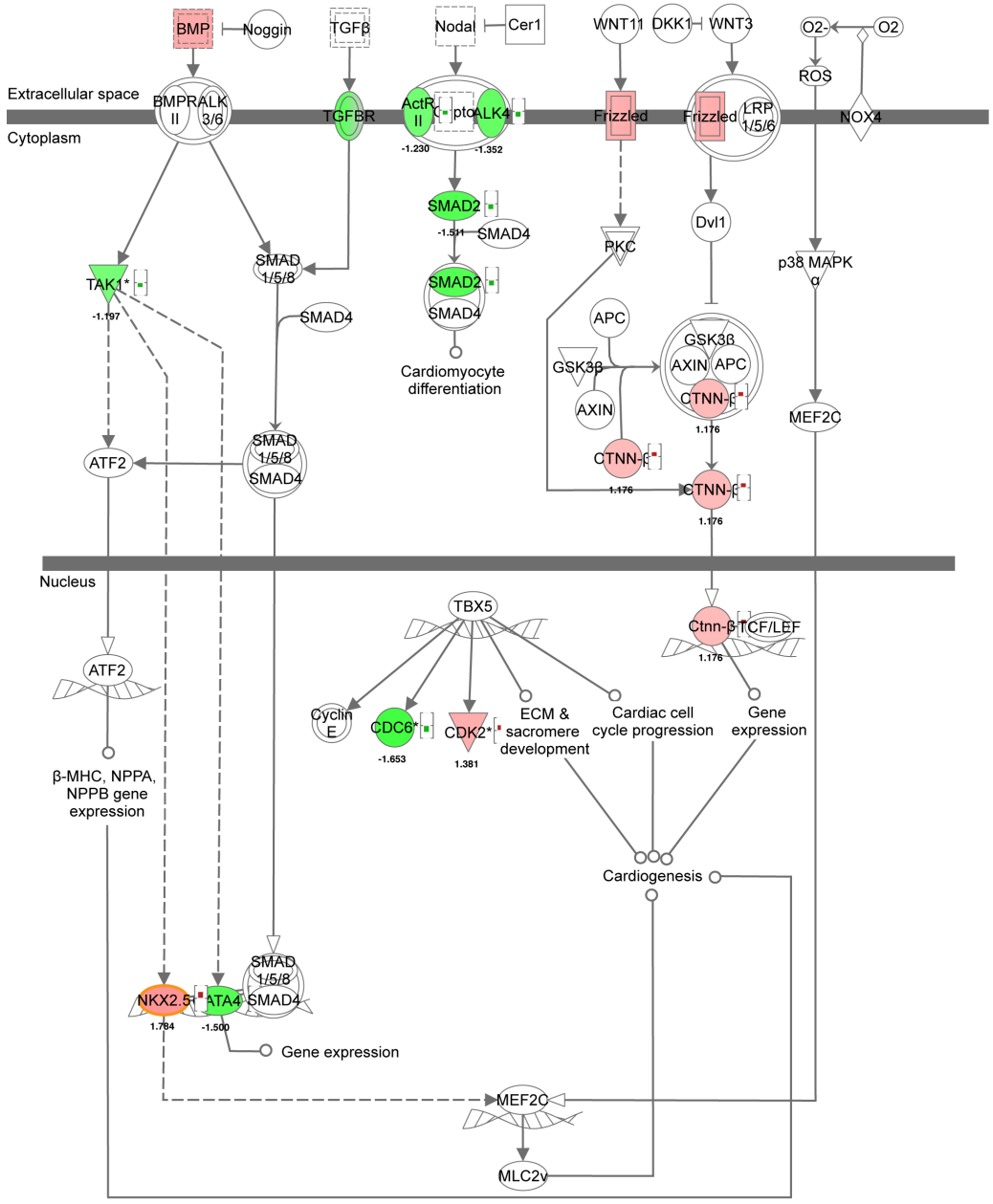


Figure 2.11 – IPA Canonical pathway: Factors Promoting Cardiogenesis in Vertebrates

Building upon the earlier Ingenuity generated pathways using *Mus musculus* orthologs to *Xenopus laevis* genes, canonical pathways associated with Nkx2-5 were chosen based upon how many members were differentially expressed. Nodes are colored based upon their fold change. Genes enriched in the Nkx2-5 overexpression samples are shown in red and those enriched in the GFP control samples are shown in green.

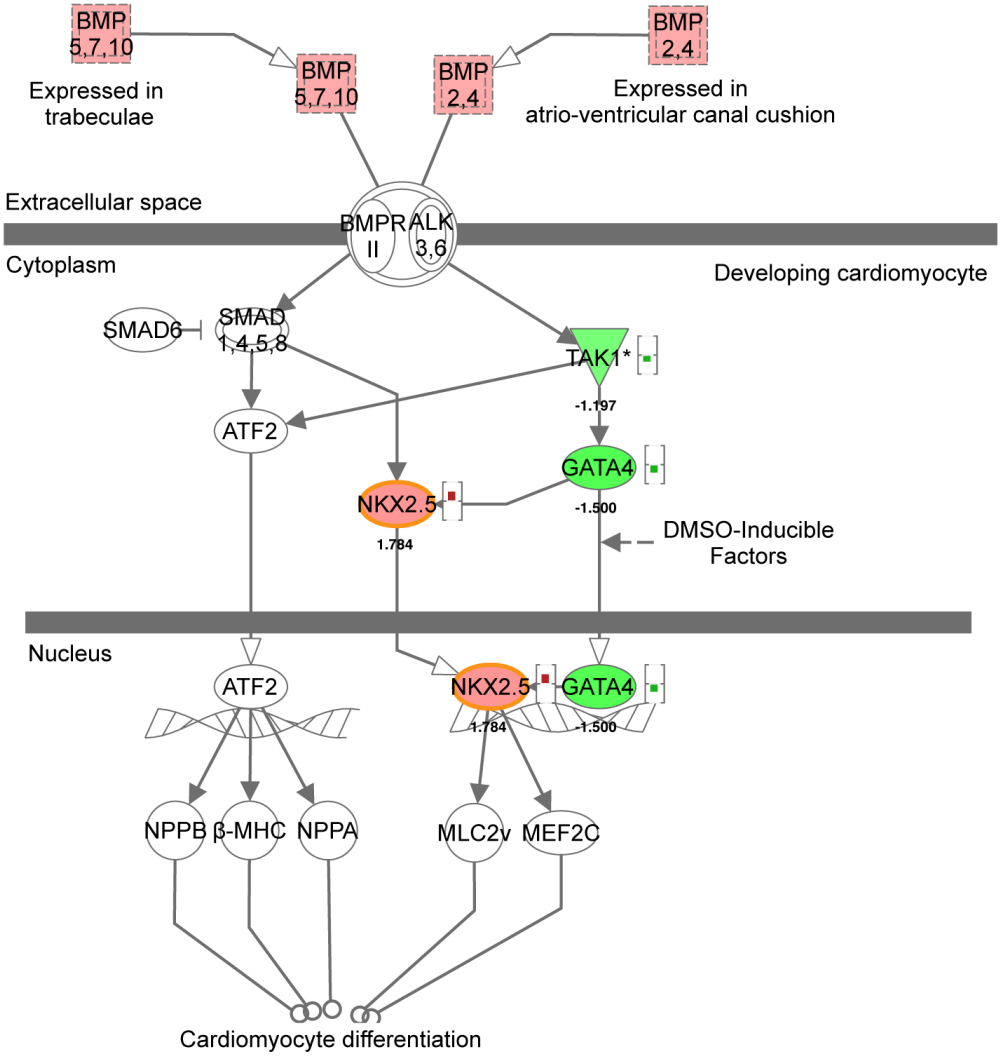


Figure 2.12 – IPA Canonical pathway: Cardiomyocyte Differentiation via BMP Receptors

Building upon the earlier Ingenuity generated pathways using *Mus musculus* orthologs to *Xenopus laevis* genes, canonical pathways associated with Nkx2-5 were chosen based upon how many members were differentially expressed. Nodes are colored based upon their fold change. Genes enriched in the Nkx2-5 overexpression samples are shown in red and those enriched in the GFP control samples are shown in green.

the head samples; 263 probesets were classified as "head-enriched" (Table 2.1), among which 24 were in the list of potential Nkx2-5 targets (Table 2.7).

Heart and transcription-related classification

Based on gene ontology annotations from the CrossGene database, probesets that were annotated with either GO:0007507 (heart development) or GO:0003015 (heart process) were classified as "heart related". In the set of 710 potential targets, 17 were classified as heart related (Table 2.7). Probesets annotated with GO:0003700 (transcription factor activity) or GO:0045449 (regulation of transcription) were classified as "transcription related" yielding 79 potential targets (Table 2.7).

A smaller prioritized list of potential targets was created based upon these three criteria: enriched expression in the head-region, heart-related GO annotation, and transcription-related GO annotation. The union of these classifications yielded 99 unique UniGene clusters, representing the most likely potential targets of Nkx2-5 related to heart development (Figure 2.15).

Presence of possible Nkx2-5 binding sites

Nkx2-5 has two binding sites: an NK2 site and a homeobox (HBOX) site (Chen et al. 1995). An additional method for prioritizing potential targets for future study is to look for Nkx2-5 binding sites in the promoters of those targets. As a substitute for the *Xenopus laevis* genome, the *Xenopus tropicalis* 4.1 assembly was used. Predicted promoter regions were searched for potential Nkx2-5 binding sites. Of the 99 prioritized targets, 90 contained at least one HBOX site and 81 contained at least one NK2 site (Table 2.7).

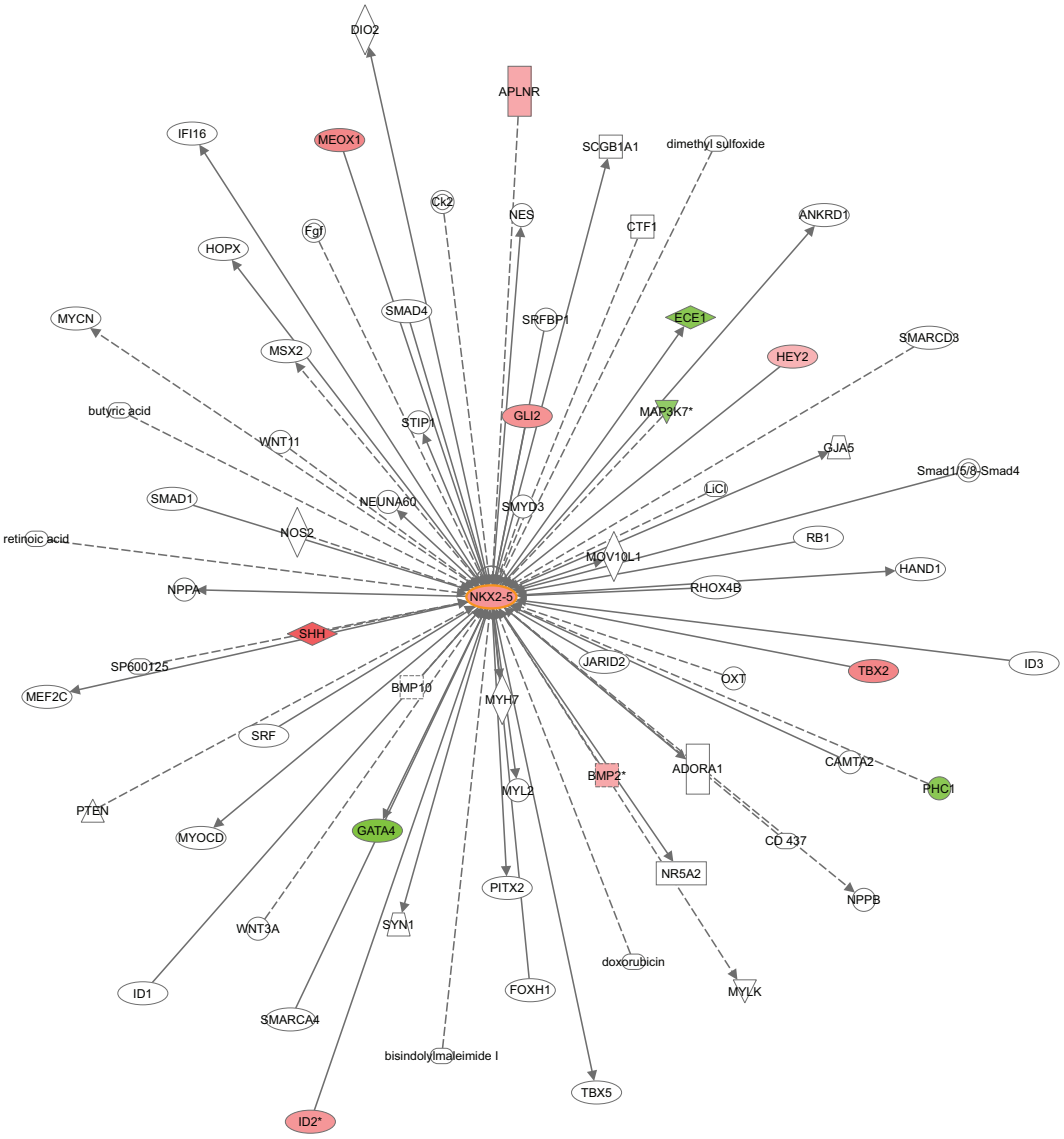


Figure 2.13 – Known Nkx2-5 interacting partners

This shows a network of all the genes and molecules in Ingenuity's Knowledge Base known to interact with Nkx2-5 (center). 12 out of 69 interacting partners showed differences in gene expression in response to Nkx2-5 overexpression in stage 11.5 *Xenopus* embryos. Direct interactions are shown in solid lines and indirect interactions are shown in dashed lines. Genes that were up-regulated in Nkx2-5 injected samples are shown in red. Down-regulated genes are shown in green.

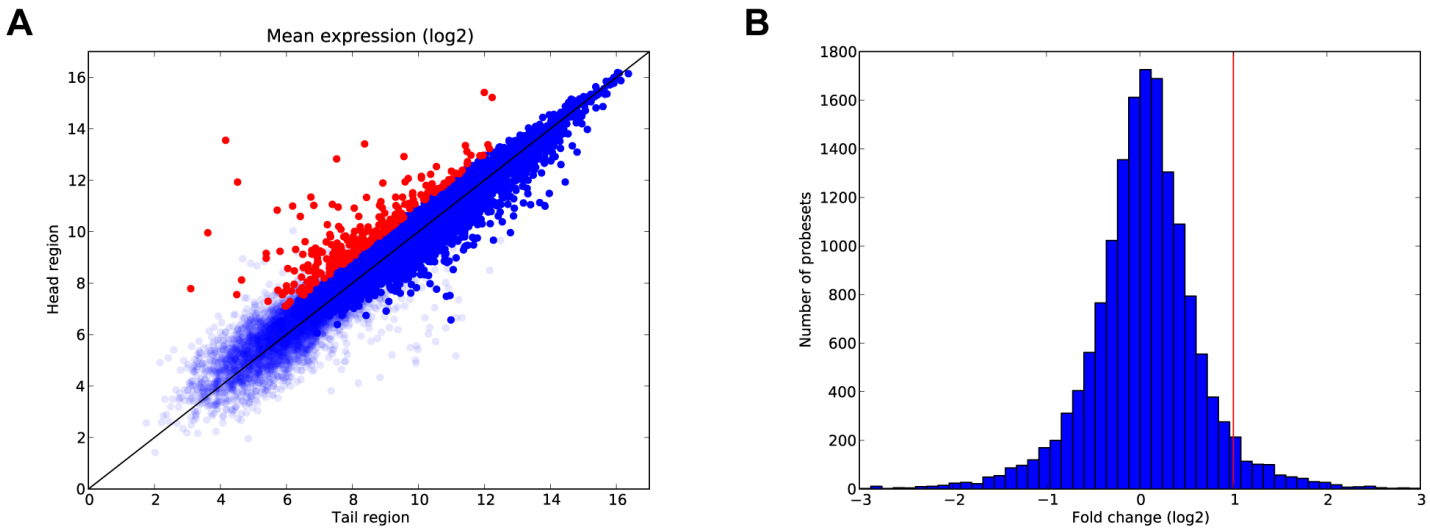


Figure 2.14 – Microarray results head versus tail samples

A) Mean expression levels (log_2 transformed) for each probeset in the head versus tail samples. Probesets with a fold change of >1.5X in the head samples are in red. B) Fold-change (log2 transformed) histogram in head versus tail samples. The vertical red line depicts the +1.5X cutoff value.

Table 2.7 – Prioritized list of potential targets of Nkx2-5

					Classification			TFBS			
UniGene	**Gene**	**Probeset**	**Fold**	**FDR**	**Heart**	**Head**	**TF**	**HBOX**	**NK2**	**NK2-HBOX**	**NK2-NK2**
Xl.879	twist1-A	Xl.879.1.S1_s_at	3.0	0.06	X	X	X	X	X	X	
Xl.22859	nkx2.5	Xl.1093.1.S1_at	1.8	0.04	X	X	X	X	X		
Xl.49495	Pax3	Xl.25932.2.A1_s_at	2.4	0.07	X	X	X	X	X	X	
Xl.277	Xhox3	Xl.277.1.S1_at	1.8	0.06	X		X	X	X	X	
Xl.508	fd-4'	Xl.508.1.S1_at	1.7	0.08	X		X	X	X	X	
Xl.866	SHH*	Xl.866.1.S2_at	2.8	0.05	X		X	X	X	X	
Xl.933	bra3-a	Xl.933.1.S1_at	1.8	0.03	X		X	X			
Xl.975	Tbx3	Xl.975.1.S1_at	2.0	0.03	X		X	X	X	X	
Xl.1042	Xgli4	Xl.481.1.S1_at	1.8	0.04	X		X	X	X	X	
Xl.8624	idb4	Xl.8624.1.S2_at	1.8	0.04	X		X	X	X	X	
Xl.25598	CITED2*	Xl.25598.1.S1_at	-1.9	0.03	X		X	X	X	X	
Xl.47015	Mad2	Xl.85.1.S1_a_at	-1.5	0.03	X		X	X	X	X	
Xl.72728	XHOXC8*	Xl.16273.1.A1_at	2.8	0.08	X		X	X	X	X	

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.708	lft-a	Xl.708.1.S1_at	-1.7	0.07	X						
Xl.951	itga3-A	Xl.951.1.S1_at	1.8	0.04	X			X			
Xl.1115	acta1	AFFX-Xl-a1Act-3_s_at	7.8	0.04	X			X	X		
Xl.15645	DSC3*	Xl.15645.1.S1_at	-1.7	0.06	X			X	X	X	X
Xl.627	six3-A	Xl.627.1.S1_at	1.6	0.06		X	X	X			
Xl.1043	zic5-A	Xl.1043.1.S1_at	1.7	0.05		X	X	X	X		X
Xl.23512	hoxa1-A	Xl.23512.1.S1_at	1.8	0.07		X	X	X	X		
Xl.28002	XGATA-3	Xl.793.1.A1_at	2.1	0.02		X	X	X	X	X	
Xl.64079	MCM5*	Xl.14322.1.A1_at	2.1	0.06		X	X	X	X		
Xl.1589	AGR3*	Xl.1589.1.S1_at	1.7	0.09		X		X	X		X
(Xl.1589)	(AGR3)	Xl.1589.1.S2_at	1.7	0.10							
Xl.1685	AGR2*	Xl.1685.1.S1_at	2.6	0.04		X		X	X		

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
(Xl.1685)	(AGR2)	Xl.25847.1.S1_at	2.2	0.06							
Xl.6393	CDC45L*	Xl.6393.1.S1_at	-2.0	0.04		X		X	X		
Xl.8333	KRCP*	Xl.8333.1.A1_at	-2.1	0.01		X		X	X		X
Xl.8926	GALNTL1*	Xl.8926.1.A1_at	1.9	0.03		X		X	X	X	
Xl.10868	xgalectin-VIa	Xl.10868.1.S1_at	2.0	0.08		X		X	X	X	
Xl.12160	tnrc4	Xl.12160.1.S1_at	2.3	0.04		X		X	X		
Xl.12275	EVI1*	Xl.12275.1.S1_at	2.6	0.06		X					
Xl.14675		Xl.15930.1.A1_at	2.0	0.04		X					
Xl.15242	STAF*	Xl.15242.1.A1_at	-1.7	0.03		X					
Xl.16169	CDH11*	Xl.16169.1.S1_at	1.8	0.04		X		X	X	X	
Xl.21346	rab3a	Xl.21346.1.A1_at	-1.9	0.09		X		X	X	X	
Xl.51651		Xl.13885.1.A1_s_at	1.7	0.05		X		X	X	X	

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.56738	chn1	Xl.9113.1.A1_at	2.1	0.02		X		X	X	X	
Xl.80278		Xl.15793.1.A1_at	1.9	0.04		X					
Xl.80502		Xl.14883.1.A1_at	-1.7	0.06		X		X	X	X	
Xl.55	HEN2-A*	Xl.55.1.S1_at	1.7	0.06			X	X	X	X	
Xl.146	XMyf-5	Xl.146.1.S1_at	2.2	0.03			X	X	X		
Xl.283	HoxA1	Xl.283.1.S1_at	2.0	0.06			X	X	X		
Xl.397	bix2-A	Xl.397.1.S1_at	-1.9	0.05			X	X	X		
Xl.475	C3H-2	Xl.475.1.S1_s_at	1.8	0.02			X	X	X		X
Xl.483	Mespo	Xl.483.1.S1_at	2.0	0.05			X	X	X	X	X
Xl.541	jun	Xl.541.1.S1_s_at	2.5	0.04			X	X	X	X	
Xl.586	ESR-7	Xl.586.1.S1_at	-2.0	0.09			X	X	X	X	
Xl.769	ets2a-A	Xl.769.1.S1_at	-2.1	0.01			X				

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.792	XGATA-2	Xl.6054.1.A1_at	1.5	0.08			X	X	X		
(Xl.792)	(XGATA-2)	Xl.792.1.S1_at	1.5	0.06							
Xl.823	mf25	Xl.823.1.S1_at	1.9	0.04			X	X	X		
Xl.824	Mix.1	Xl.824.1.S1_at	-2.4	0.03			X	X	X	X	
Xl.847	OTOG*	Xl.5324.1.S1_at	3.4	0.01			X				
Xl.1130	pax2-A	Xl.1130.2.S1_a_at	1.7	0.07			X	X	X	X	
Xl.1201	xGR	Xl.21632.1.S1_s_at	-2.0	0.08			X	X	X	X	
Xl.1209	XlHbox1	Xl.1209.1.S1_at	2.4	0.04			X	X	X	X	
Xl.1249	myod1-a	Xl.1249.2.S1_a_at	1.6	0.06			X	X	X		
Xl.1419	IRF1*	Xl.1419.1.A1_at	1.8	0.06			X	X	X	X	
Xl.1502	meox2	Xl.1502.1.S1_at	2.0	0.10			X	X			
Xl.2518	SETD8-A*	Xl.2518.1.S1_at	-1.5	0.02			X	X	X		

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.4276	CCNA1*	Xl.4276.1.S1_at	-1.7	0.07			X	X	X	X	
Xl.4665	BRF2*	Xl.4665.1.A1_at	-1.5	0.05			X	X	X	X	
Xl.4898	ing3	Xl.4898.1.S1_at	-1.6	0.04			X	X	X	X	
Xl.5240	CEBPG*	Xl.5240.1.A1_at	-2.4	0.04			X	X			
Xl.5284	B4GALT6*	Xl.9728.1.A1_at	1.7	0.07			X	X	X	X	
Xl.6282	CCNA1*	Xl.6282.1.S1_at	-1.7	0.07			X	X	X	X	
Xl.6573	hbox10-A	Xl.6573.1.S1_at	-1.5	0.02			X	X	X		
Xl.7195	mcm6	Xl.7195.2.S1_at	1.7	0.10			X	X	X		
Xl.7451	WBSCR11	Xl.7451.1.S1_at	-1.7	0.04			X	X	X	X	
Xl.8168	ZBTB34*	Xl.8168.1.S1_at	-1.6	0.04			X	X	X	X	
Xl.9206	Stripy	Xl.9206.1.S1_at	1.6	0.09			X	X	X		
Xl.9271	esr10-A	Xl.9271.1.S1_at	2.6	0.01			X	X	X	X	

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
(Xl.9271)	(esr10-A)	Xl.9271.1.S1_x_at	2.1	0.05							
Xl.11454	IKBKG*	Xl.11454.1.A1_at	-1.5	0.07			X	X	X	X	
Xl.12327	EPAS1*	Xl.15970.1.A1_at	1.5	0.05			X	X	X	X	
Xl.13960	UBA3*	Xl.13960.1.A1_at	-1.9	0.04			X	X			
Xl.14312	MAZ*	Xl.14312.1.S1_at	-1.6	0.05			X	X	X	X	
Xl.16180	TBX6*	Xl.16180.1.A1_at	1.6	0.03			X				
Xl.16504	PRDM4*	Xl.16504.1.A1_at	1.8	0.03			X	X	X	X	
Xl.18653	KEAP1*	Xl.18653.1.A1_at	-1.7	0.06			X	X			
Xl.18750	NFE2L2*	Xl.18750.1.S1_at	2.0	0.02			X	X	X		
Xl.19790	IRF1*	Xl.19790.1.S1_at	1.8	0.03			X	X	X	X	
Xl.21223	ZBTB5*	Xl.4309.1.A1_at	1.5	0.01			X	X	X	X	
Xl.21817	HES1*	Xl.972.1.S1_at	1.8	0.03			X	X	X		

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.21932	etv3	Xl.21932.1.S1_at	-1.8	0.05			X	X	X		
Xl.22568	xlcry-dash	Xl.22568.1.S1_at	-1.9	0.04			X	X	X	X	
Xl.23822	staf	Xl.23822.1.S1_at	-1.6	0.03			X	X	X	X	
Xl.26888	cdx4	Xl.10269.1.S1_at	1.7	0.06			X	X			
Xl.29033	esr9	Xl.12444.1.S1_at	1.9	0.07			X				
Xl.29292	Larg	Xl.7724.1.A1_at	-1.5	0.05			X	X	X	X	
Xl.37147	Ywhab	Xl.21804.1.S1_at	1.7	0.06			X	X	X	X	
Xl.49528	HES2*	Xl.19075.1.A1_at	1.6	0.04			X	X	X	X	X
Xl.50380	HSF2	Xl.25289.1.A1_at	1.5	0.03			X	X	X		
Xl.50827	JUND*	Xl.23976.1.A1_at	1.7	0.06			X	X	X	X	X
Xl.51393	dmrta1	Xl.15931.1.A1_at	3.2	0.05			X	X	X		
Xl.53382	POU5F1*	Xl.869.1.S2_at	-1.8	0.05			X	X	X	X	

Table 2.7 (continued)

UniGene	Gene	Probeset	Fold	FDR	Classification			TFBS			
					Heart	Head	TF	HBOX	NK2	NK2-HBOX	NK2-NK2
Xl.53491	HoxD1	Xl.3370.1.S1_at	1.5	0.07			X	X	X	X	
Xl.56499	NFE2L1*	Xl.9712.1.A1_at	1.8	0.06			X	X	X	X	
Xl.57198	OCRL*	Xl.5376.1.A1_at	-1.5	0.03			X	X	X	X	
Xl.73183	MYOCD*	Xl.16301.1.A1_at	2.2	0.07			X	X			
Xl.76016	FHL3*	Xl.22844.2.A1_at	1.7	0.06			X	X	X	X	
Xl.79342	FST*	Xl.1094.1.S1_at	2.0	0.08			X	X	X	X	X

* These genes were unnamed in UniGene, the name shown was predicted using CrossGene

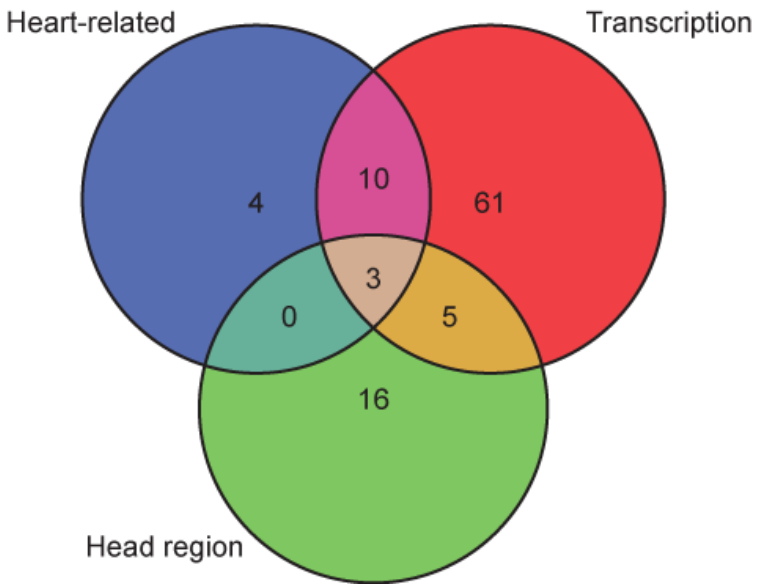


Figure 2.15 – Venn diagram showing the number of genes matching each classification type

Based upon the 709 differentially expressed genes, genes were further classified by assigning them to three overlapping groups: heart-related genes (17), transcription-related genes (79), and genes that were enriched in the head-region of the embryo (24). Three genes matched all three groups: Nkx2-5, Pax3, and Twist1.

Nkx2-5 binds DNA as a dimer, so potential binding sites near another site are particularly interesting (Kasahara et al. 2001). Additionally, in the promoter region of the known Nk2-5 target, ANF, there is an NK2 site paired with an HBOX site, which forms an enhanced NK (NKE) binding site (Small et al. 2003). Because of this, paired binding sites within 20 bases of each other were tallied. 55 of the 99 potential UniGene targets contained a NK2 site near an HBOX site (NKE), and 9 contained an NK2 site near another NK2 site (Table 2.7). 12 of the 17 heart-related targets, 10 of the 24 head-enriched targets, and 47 of the 55 transcription related targets had at least one NKE site.

Discussion

Nkx2-5 is an essential transcription factor in cardiogenesis, but overexpression of Nkx2-5 in stage 11.5 embryos showed a broader effect than previously imagined. Injection of synthetic mRNA for Nkx2-5 in the 8-cell stage led to changes in expression of thousands of genes at stage 11.5 (Table 2.1). Analyses of the pathways and networks affected showed up-regulation of transcription-factor and DNA-binding activity and developmental pathways, including cardiogenesis. The GO enrichment results also suggest that developmental genes are primarily up-regulated by over-expression of Nkx2-5. Using a combination of regional expression and transcription- or heart-related annotation, a list of 99 prioritized targets was compiled (Table 2.7). The expression changes could be due to direct or indirect effects; the presence of putative Nkx2-5 binding sites in the promoters of many of the genes suggests direct effects on many of the genes.

Two of the more interesting potential targets are Twist1 and Pax3. In the CrossGene database, Twist1 and Pax3 were both annotated as heart and transcription related, and they were both enriched in the head regions. The only other gene to have this annotation pattern was Nkx2-5 itself, which is known to be auto-regulatory (Oka et al. 1997; Reecy et al. 1999); this finding serves as a positive control for our approach. The synthetic Nkx2-5 mRNA that was injected into the embryos had a different 3' UTR from endogenous Nkx2-5. Because of this, the GeneChip probesets for Nkx2-5 did not detect the injected Nkx2-5HA mRNA. Instead, the an increase in the expression of endogenous Nkx2-5 was detected, confirming the auto-regulation of Nkx2-5. Twist1 and Pax3 both contain paired Nk2/HBOX binding sites. In *Drosophila melanogaster, twist* directly regulates the Nkx2-5 ortholog *tinman* (Bodmer 1993; Yin et al. 1997) and it has been shown by a loss-of-function study to have an effect on cardiac neural crest cells in mouse (Vincentz et al. 2008). In *Xenopus laevis,* it was reported that they could not directly regulate each other during early development, as their expression patterns did not overlap at the appropriate stages (Evans et al. 1995). However, at earlier stages of development, their expression patterns may overlap enough to allow some degree of regulation. In *Xenopus laevis,* the Pax3 neurula expression pattern seems to be restricted to the anterior neural fold and doesn't appear to overlap with Nkx2-5 (Xenbase image: 20246)(Bowes et al. 2010). However, in mouse, Pax3 is essential for heart development and is required for cardiac neural crest migration (Conway et al. 1997a; Conway et al. 1997b). This suggests a potential for interactions later in development.

These results are the first step in the process of uncovering the earliest role and targets of Nkx2-5 in the developing embryo. It also represents a strategy for analysis of future

genome scale experiments in *Xenopus laevis*. With the help of cross-species annotations using CrossGene, it was shown that the overexpression of Nkx2-5 had large effects on development and transcription related pathways. Finally, on the basis of the changes in gene expression, external annotations, gene expression data from a bisected embryo, and the surrogate *Xenopus tropicalis* genome, a list of potential targets was compiled. These potential targets can be tested to establish a direct connection with Nkx2-5.

CHAPTER 3: EXPRESSION PROFILING OF SELECTED TARGETS

Introduction

Expression profiling of a gene involves measuring the amount of a specific RNA molecule in a sample over various conditions. This establishes a pattern of expression for that gene. If two genes have similar expression patterns, this indicates that they could share common regulatory signals. If one of those genes is auto-regulatory, such as Nkx2-5, it could mean that the shared regulation between two genes may be the auto-regulatory gene itself. In general, a gene that auto-regulates itself will increase in abundance until it reaches a steady-state level (Figure 3.1A). However, if expression of the gene is restricted to a subset of cells and this subset of cells makes up a decreasing percentage of the developing organism, the observed abundance of the auto-regulatory gene will show a different expression pattern (Figure 3.1B). While correlation of expression patterns does not imply a causal relationship, it is a useful tool for gauging the degree to which genes could be co-regulated. Commonly used techniques for profiling gene expression includes two PCR methods: semi-quantitative RT-PCR and quantitative real-time PCR.

Semi-quantitative RT-PCR profiling

With semi-quantitative RT-PCR profiling, RNA from a sample is extracted, reverse transcribed to cDNA, and then the cDNA is used as a template for PCR amplification. This doesn't use any special dyes or equipment and is visualized on a standard agarose gel with ethidium bromide (EtBr) staining. In most cases, the goal with PCR is to generate as much of the target DNA as possible. However, the geometric expansion of

PCR produces so much DNA that when the bands are visualized, the image may be saturated, showing only whether or not a band is present, but making it difficult to see if one sample had more or less DNA originally. With semi-quantitative RT-PCR, the goal is to generate an amount of DNA that falls below the saturation point of the PCR conditions, but enough to still allow the visualization of a band. This way is it possible to determine the relative amount of DNA in one sample compared to another. One method for adjusting the amount of DNA produced with PCR is controlling the number of cycles of PCR amplification. A common number of cycles for saturating PCR amplification to completion is 40. In semi-quantitative RT-PCR, this may be adjusted to be between 25-35. This doesn't produce an absolute number, but it does allow for rough quantification of one gene relative to another. Because it doesn't require any extra equipment or dyes, it is also a great deal cheaper than more quantitative methods.

Quantitative real-time PCR

Quantitative real-time PCR (qPCR), on the other hand, does allow for the absolute quantification of an RNA molecule present in a sample (Heid et al. 1996). With qPCR, like RT-PCR, RNA is extracted from a sample and reverse transcribed to cDNA. This cDNA is then used as a template in PCR amplification. With qPCR, however, this is done in the presence of special dyes that measure the quantity of DNA present for each cycle of the PCR amplification. In order for the quantity to be measured after each cycle, the reaction takes place in a special real-time PCR machine that includes optics necessary to measure the fluorescence of the dyes (Wittwer et al. 1997). The amount of fluorescence of the dyes corresponds to the amount of DNA present in the reaction. With each cycle of PCR amplification, more and more DNA is produced, which causes an increase in

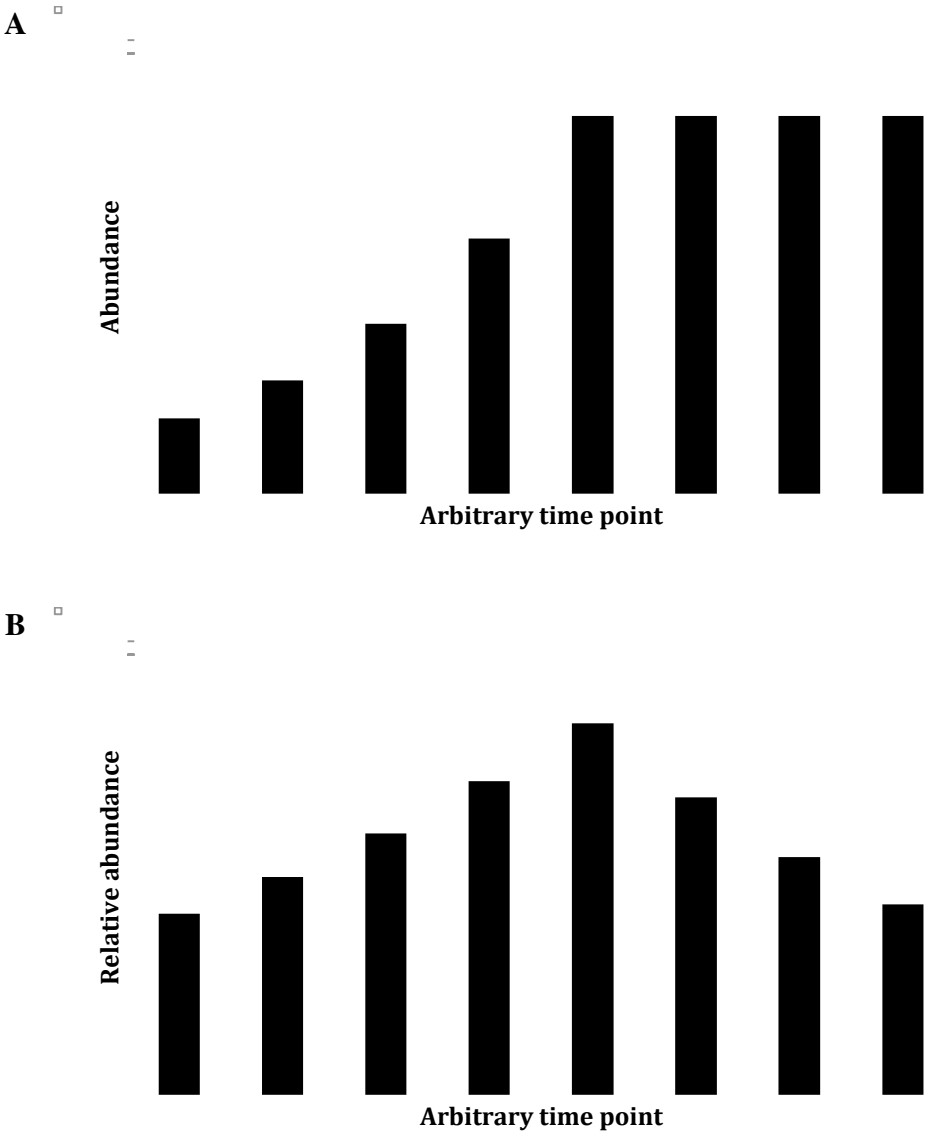


Figure 3.1 – Model of gene expression for an auto-regulatory gene in a developing organism

This is a predicted model of gene expression for an auto-regulatory gene. A) In this model, the expression of a gene is only dependent upon the abundance of a gene in the prior time point. The abundance of the gene eventually plateaus to a constant level. The gene is assumed to be expressed ubiquitously in the sample. B) In this model, the expression of a gene is still dependent on the abundance of a gene in the prior time point. However, this adds the assumption that the gene of interest is only expressed in a particular cell population that makes up a smaller and smaller percentage of the organism as time progresses. In this model, the abundance increases until the plateau stage is reached. After this point, the abundance appears to decrease, relative to the size of the organism.

fluorescence. There are a number of dyes used in qPCR, but one of the most common is quantification using the commercially available dye SYBR Green (Invitrogen) (Schneeberger et al. 1995; Wittwer et al. 1997). SYBR Green preferentially binds double stranded DNA over single-stranded DNA, which makes it a good choice for detecting the presence of double-stranded PCR products.

The abundance of DNA in a sample can be quantified using the C_t method (Livak 1997). When fluorescence is plotted relative to the PCR cycle, there is a region during each reaction when amplification isn't limited by reagent concentrations and fluorescence increases exponentially. When plotted on a logarithmic scale, this is called the log-linear region. A threshold line can then be set that represents a constant amount of fluorescence across all samples. This line is usually set within the log-linear region. When the intersection of this line with the amplification plot is projected onto the cycle axis, the threshold cycle (C_t) is determined. This doesn't need to correspond to a discrete cycle, and can be a partial cycle. The C_t value is the theoretic cycle where all samples have the same amount of DNA present. When combined with a standard curve, this value makes it possible to calculate how much DNA was present in the initial preparation (C_0).

Based on the initial microarray results (Chapter 2), and an early version of the CrossGene database (Chapter 4), 34 genes were selected for profiling with semi-quantitative RT-PCR. These experiments were performed using a prototype of the CrossGene database (Chapter 4) that wasn't as comprehensive or complete. Because of this, candidate genes were selected using a slightly different methodology than that described in Chapter 2. Even when using these different criteria, the 5 genes selected for qPCR profiling were also present in the list of potential targets detailed in Chapter 2: Zic5, HBox1, CHN1,

Bix2, and Mix1. The expression pattern of these genes was then compared with the expression pattern of Nkx2-5 to determine the correlation.

Methods

Candidate gene selection

Using the microarray results presented in Chapter 2, candidate genes were selected for further analysis using a combination of statistical and annotation criteria (Table 3.1). Probe sets were assigned a *Xenopus laevis* UniGene cluster ID by comparing the probe set's target sequence to *Xenopus laevis* UniGene build 66 using the BLASTN algorithm (Altschul et al. 1990a; Altschul et al. 1997a). After statistical filtering, candidate genes were filtered by the presence of a consensus Nkx2-5 binding site ([TC]AAGTG) in the corresponding *Xenopus tropicalis* promoter (up to 2,000 bp upstream of the annotated transcriptional start site). Because CrossGene was not yet available, instead of using the CrossGene reciprocal groups to find orthologs, only orthologs from *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus* were considered. Orthologs were found by TBLASTX comparison of the representative *Xenopus laevis* UniGene cluster sequence for a probe set to the UniGene clusters for the other organism. The single best match was taken as the ortholog, regardless of reciprocity. GO annotations for the orthologous genes were obtained from the NCBI Gene database (Maglott et al. 2007). For each *Xenopus laevis* gene, if any of the orthologs were annotated with the GO terms listed in Table 3.2, the *Xenopus laevis* gene was considered to share that annotation.

Table 3.1 – Selection criteria for candidate genes

These are the selection criteria used to determine which candidate genes were to be profiled using semi-quantitative RT-PCR. First genes had to be called "present" in at least 5 out of 6 Nkx2-5 injected samples. Next, genes needed to have an absolute fold change greater than 1.5X and a FDR less than or equal to 0.1. Then, genes without a consensus Nkx2-5 binding site in the promoter of the gene were removed (using *Xenopus tropicalis* promoters 2,000 bp upstream of the predicted start site). This resulted in 318 genes. From these, the only genes that were kept were a) annotated with a heart related GO term (Table 3.2), b) enriched in Nkx2-5 samples and were annotated as having transcription factor activity (Table 3.2), or c) predominantly in the "head" region (Table 2.1). This resulted in 34 candidate genes.

Criteria		Matching
No filter		15,503
Present in Nkx2-5 samples (Absent/Present filter)	≥0.83	10,402
Fold change	±1.5X	826
FDR q-value	≤0.1	509
Consensus Nkx2-5 binding site present in 2,000 bp upstream		318
a) GO: Heart		10
b) GO: Transcription factor and enriched in Nkx2-5 samples		17
c) Predominantly in "head" samples		11
Unique candidate genes		34

Table 3.2 – GO terms used for candidate gene selection

The human, mouse, or rat homologs of a *Xenopus laevis* transcript had to be annotated with one of these GO terms in order to be considered for further testing. For this selection, all of the first four GO terms were all considered "Heart"-related.

Description	GO Term
Heart development	GO:0007507
Embryonic heart tube development	GO:0035050
Heart morphogenesis	GO:0003007
Regulation of heart contraction	GO:0008016
Transcription factor activity	GO:0003700

Semi-quantitative RT-PCR profiling

Primers for semi-quantitative RT-PCR profiling were designed for each gene algorithmically using Primer3 software with the parameters listed in table A1.1 (Rozen et al. 2000). The target product size was 1,000 bp using the 3' half of the UniGene sequence as the template. Total RNAs were collected as previously described from pools of *Xenopus laevis* embryos at the following stages and adult tissues: Unfertilized egg, St. 9, St. 10.5, St. 11, St. 11.5, St. 12, St. 14, St. 16, St. 18, St. 21, St. 26, St. 28, St. 30, St. 32, St. 35, spleen, liver, heart, and skeletal muscle. 10 μg of each RNA sample was reverse transcribed into cDNA using a SuperScript II Reverse Transcriptase kit (Invitrogen) in a 40 μl reaction using anchored oligo-dT as a primer. The resulting cDNAs were diluted 12.5X to 500 μl yielding a concentration of 20 ng/μl. Next, for each gene, 5 μl of the diluted cDNAs were amplified by PCR using Platinum *Taq* DNA polymerase (Invitrogen) using the following PCR conditions: 94°C for 2 minutes; 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds, 72°C for 75 seconds; followed by 72°C for 2 minutes. PCR products were visualized with agarose gel electrophoresis and EtBr staining. ODC (NCBI Gene ID: 379859) and β-actin (NCBI Gene ID: 398459) were used as internal controls.

Quantitative real-time PCR profiling

Primer design

Since quantitative real-time PCR is very sensitive to genomic DNA contamination, it is advantageous to design PCR primers that span an exon/exon boundary. As previously described in Chapter 2, genomic assemblies for *Xenopus tropicalis* were used as an

analog for *Xenopus laevis* genomic sequence. The annotations for *Xenopus tropicalis* predicted transcripts also include predicted exon locations. For CHN1, Zic5, HBox1, Bix2, GAPDH, and Nkx2-5, new primer pairs were designed using Primer3 software (Rozen et al. 2000) with a target PCR product size of 150-225 bp. Additionally, for all genes except Nkx2-5, the primers were designed to include the last predicted exon/exon boundary. For Nkx2-5, the primers were designed against the full length of gene. Then, using custom written software, potential PCR products were predicted for each primer pair using all *Xenopus laevis* UniGene clusters as the template. Up to two mismatches were allowed, to allow for flexibility in the PCR product predictions. Only primer pairs that were predicted to have a single product were selected. For ODC and Mix1, this strategy failed to produce confirmable PCR products, so previously published qPCR primers were used instead (Heasman et al. 2000; Xanthos et al. 2001).

Cloning control PCR fragments

Each primer pair was confirmed by PCR to produce a single band of the appropriate size using either pooled St. 10.5 cDNA or St. 45 cDNA. For PCR amplification, recombinant *Taq* polymerase (Invitrogen) was used with the following PCR reaction conditions: 94°C for 3 minutes, followed by 40 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds. PCR products were visualized with agarose gel electrophoresis with EtBr staining as described in Chapter 2.

The PCR products were then purified using a QIAquick PCR purification kit (Qiagen, Valencia, CA) and cloned into a pCRII plasmid using a TOPO TA Cloning kit (Invitrogen) and transformed into DH5α E. coli cells (Invitrogen). Colonies were then selectively grown on LB/agar plates containing 100 μg/ml of ampicillin (Sigma-Aldrich,

St. Louis, MO). Individual colonies were picked and grown in 4 ml of LB broth (Invitrogen). Plasmids were then purified using a QIAprep spin miniprep kit (Qiagen) and their inserts confirmed by PCR using gene specific and generic M13 primers. PCR was performed as described above, with the exception that with M13 primers, the extension step was extended to 45 seconds.

The cloned control PCR products were linearized by incubating 1 μg of the plasmid in a 20 μl reaction containing 2 μl 10X Buffer D (6 mM Tris-HCl, pH 7.9; 150 mM NaCl; 6 mM $MgCl_2$; 1 mM DTT) and 10 U *NotI* restriction enzyme (Promega, Madison, WI). The reaction was incubated at 37°C for one hour, followed by heat inactivation of the enzyme by incubating the mixture at 65°C for 15 minutes. Serial dilutions were then made of this reaction mixture. These dilutions served as the basis for standard curves in the subsequent quantitative real-time PCR assays so that each gene could be quantified.

RNA extraction from fixed embryos

For qPCR analysis, six embryos from each of the following stages were used: St. 11, St. 17, St. 22, St. 28, and St. 45. RNA was extracted from embryos previously fixed in NOTOXhisto (Scientific Device Laboratory, Des Plaines, Illinois) and stored at 4°C. For these samples, individual embryos were removed from methanol, placed in a microcentrifuge tube containing 100 μl of Proteinase K buffer (0.1 M NaCl, 10 mM Tris-HCl pH 8, 1 mM EDTA, 0.5% SDS), homogenized with a tube pestle, and incubated for 90 minutes with 1.5 μl Proteinase K (20 mg/ml) (Ambion) in at 50°C while being agitated. RNA was then extracted and cDNA synthesized using the same TRIzol procedure detailed in Chapter 2. After cDNA synthesis, the cDNAs were diluted to a final concentration of 5 ng/μl.

Real-time qPCR profiling

Real-time qPCR was performed using the Roche LightCycler® 480 (Roche Applied Science, Indianapolis, IN) using the Express SYBR GreenER qPCR mix with ROX (Invitrogen). All reactions used the following PCR conditions: 95°C for 5 minutes and 45 cycles of 95°C for 10 seconds, 60°C for 5 seconds, 72°C for 10 seconds. Data acquisition was performed after the 72°C extension step for all PCR cycles. After the PCR cycles were complete, melting curve analysis was performed to confirm product quality. The reaction conditions for each well were: 10 μl master mix, 0.4 μl gene-specific primers (100 pmol/μl), 5.6 μl water, and 4 μl of template (cDNA or diluted linearized plasmid).

For each gene tested, two 96 well plates were used. Only one gene was tested per plate. The first plate contained cDNAs for stages 11, 17, and 22. The second contained cDNAs for stages 28 and 45. Each stage was represented by 6 biological replicates (individual embryos) with 4 technical replicates each (24 total wells). Each plate also contained 4 non-template negative control wells and a standard curve. The standard curve was made up of seven 1:10 serial dilutions of the cloned control PCR products (from 10^{-4} to 10^{-10}) in duplicate. For these wells, the number of copies present in each well was calculated based upon the concentration of DNA, as measured by a NanoDrop spectrophotometer (Ambion), and the molecular weight of the cloned control PCR fragment (Figure 3.2).

Measuring RNA abundance

C_t values were calculated using a technique similar to that described by Ramakers, et al (Ramakers et al. 2003). For each well, background was calculated as the minimal amount of fluorescence measured for the well. Background was then subtracted from the fluorescence values and the fluorescence value was log transformed. The log-linear range

A $$mass_{copy}\frac{g}{copy} = total\ size\ (bp) \times 1.1_{10}{}^{-21}\frac{g}{bp}$$

B $$\text{copies} = \text{conc}\ \frac{\text{ng}}{\mu\text{l}} \times \frac{1}{\text{mass}_{\text{copy}}}\frac{\text{copy}}{\text{g}} \times \frac{\text{g}}{10^9\ ng} \times Volume\ \mu l$$

Figure 3.2 – Equations for calculating copy number from concentration and size of a DNA fragment

A) Calculating the mass of a linearized plasmid based on the total size of the fragment (bp) and the average mass of a fragment (1.1e-21). B) The number of copies in a sample can be calculated using the mass of the fragment and the concentration of DNA in a sample.

was found by removing the initial lag-phase and ending plateau phase PCR cycles (Figure 3.3). The optimal linear regression that encompassed three or more points in the log-linear range was then calculated (Figure 3.3). This resulted in an upper and a lower bound for the linear range. For each plate, a threshold line was picked that fell within the upper and lower bounds for the linear ranges of all samples (Figure 3.4). The threshold for each plate was determined independently. If the log-linear range for a well could not be found, started past cycle 40, or had a linear regression slope more than 4 standard deviations away from other wells, that well was discarded and not used in further analysis. The point where the threshold line crossed the log-linear range is called the C_t value and was calculated using the equation of the optimal linear regression.

The standard equations for describing PCR amplification are given in Figure 3.5 (Ramakers et al. 2003). Once C_t values were found for each well, a standard curve was calculated by plotting the C_t value of the standards versus the $\log_{10}$-transformed copy number. Next, a linear regression was calculated to find the line that best fit the data (Figure 3.6). The equation describing this line was then used to calculate the initial number of copies present at C_0 in each sample (Figure 3.5B). Mean copy number was then calculated for technical replicates. Copy number was then normalized to ODC across samples by dividing the copy number for each gene by the copy number of ODC in each sample. Normalizing helps to minimize variability when comparing one sample to another. Using the normalized copy numbers for each biological sample, correlation to the expression pattern of Nkx2-5 was then calculated using the Pearson sample correlation coefficient (Figure 3.7).

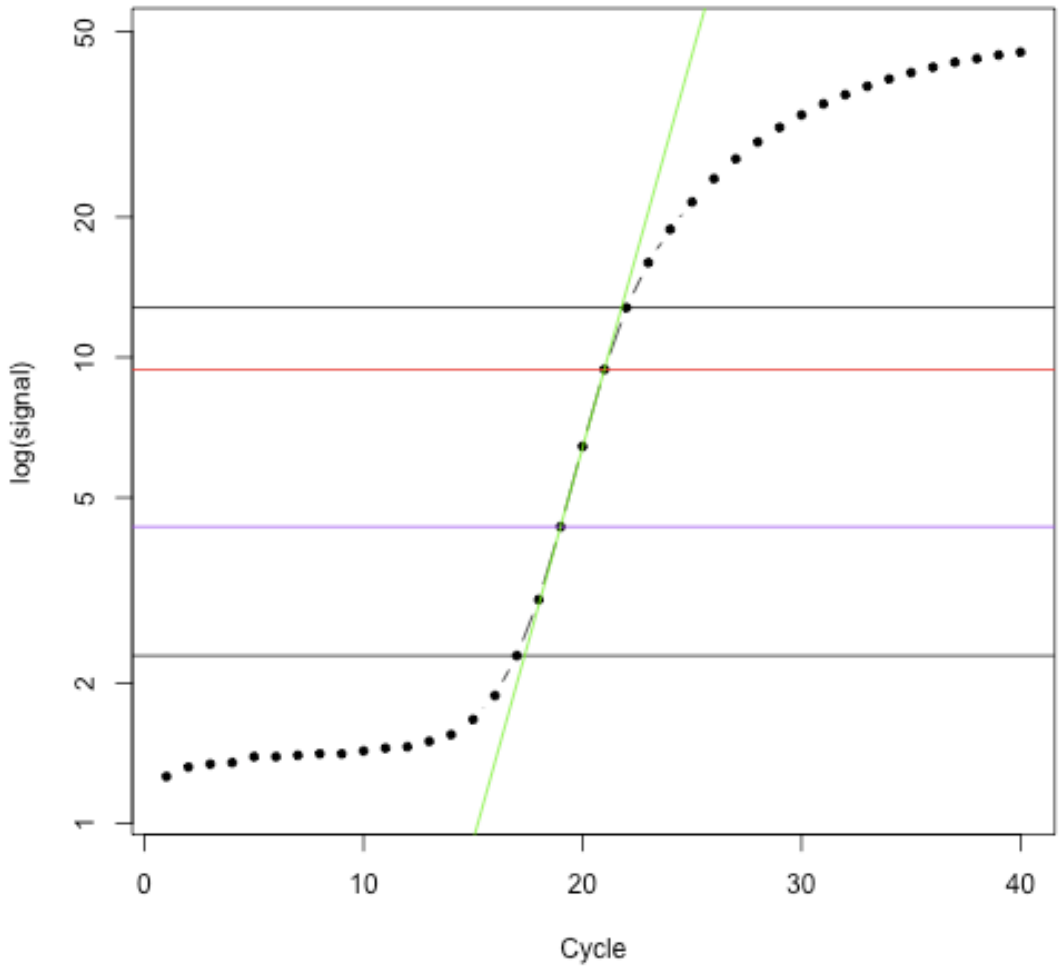


Figure 3.3 – Slope finding in a qPCR sample

In this example well, background corrected fluorescence values were plotted on a logarithmic scale versus the cycle number. An upper and lower bound (black) for testing linear regressions were calculated based on finding a starting and ending regions, representing the initial lag-phase and the non-linear ending plateau. Linear regressions were then calculated within this region for all consecutive sets of three or more cycles. The best fitting regression is then plotted (green). The upper and lower bounds of the points that make up the best-fit regression are marked in red and blue, respectively.

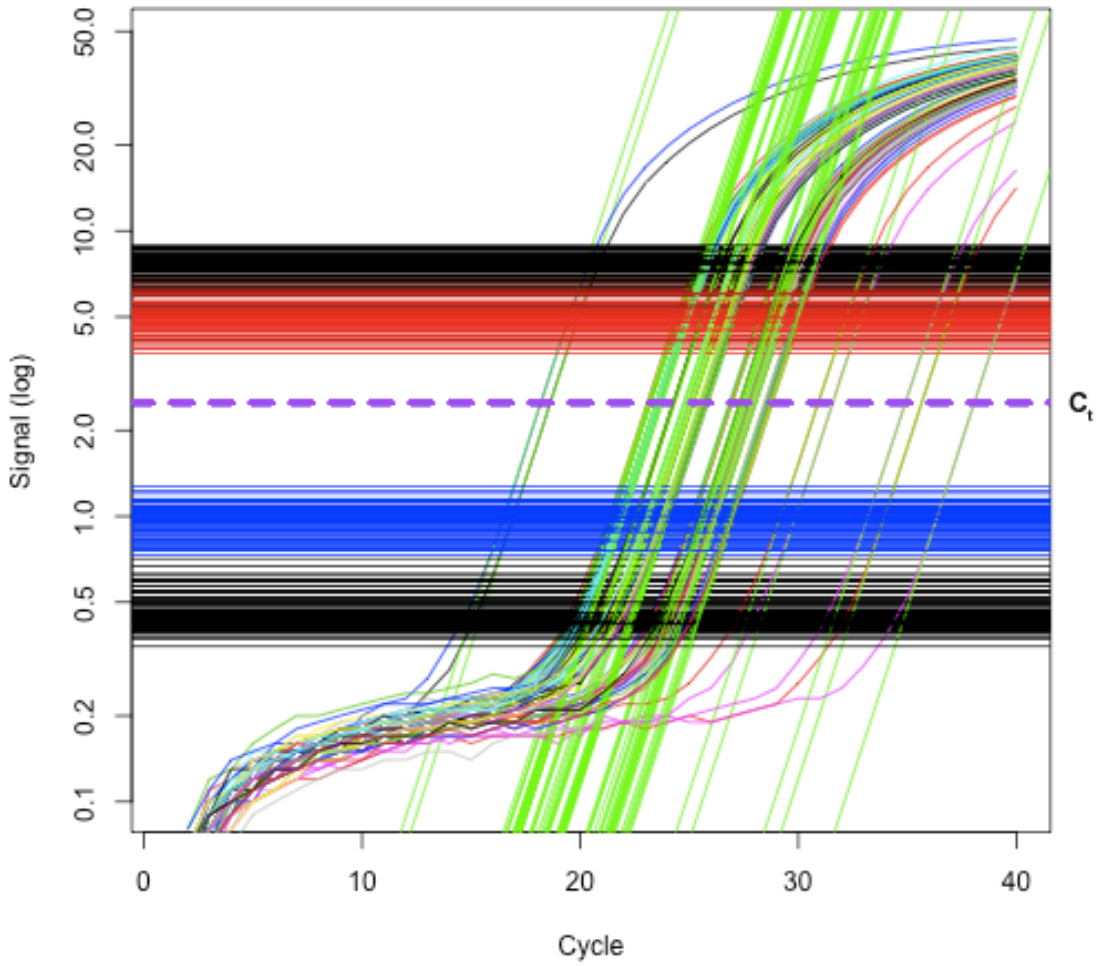


Figure 3.4 – C_t finding for a qPCR plate

The log-linear and linear regression lines plotted for an entire plate, including the standard curve. The lines and colors are as described in Figure 3.3. In this case, a threshold cycle (C_t) was determined by taking the mean between the lowest upper bound and the highest lower bound (purple). This line represents a constant amount of DNA present in all samples and is used to calculate the amount of DNA initially present in each sample.

A $$N_C = N_0 + E^C$$

B $$C = -\left(\frac{1}{\log(E)}\right)\log(N_0) + \frac{\log(N_c)}{\log(E)}$$

C $$E = 10^{-\frac{1}{slope}}$$

Figure 3.5 – Equations describing PCR amplification

A) The amount of DNA present after cycle *C* is determined by two factors – the amount of starting material (N_0) and the efficiency of the PCR reaction (E). B) Equation A, log-transformed. This shows the linear relationship between an arbitrary cycle C and the amount of starting material (N_0). If the E is constant, a linear regression can be used to determine both E and N_0. C) E can be found using the slope of the linear regression.

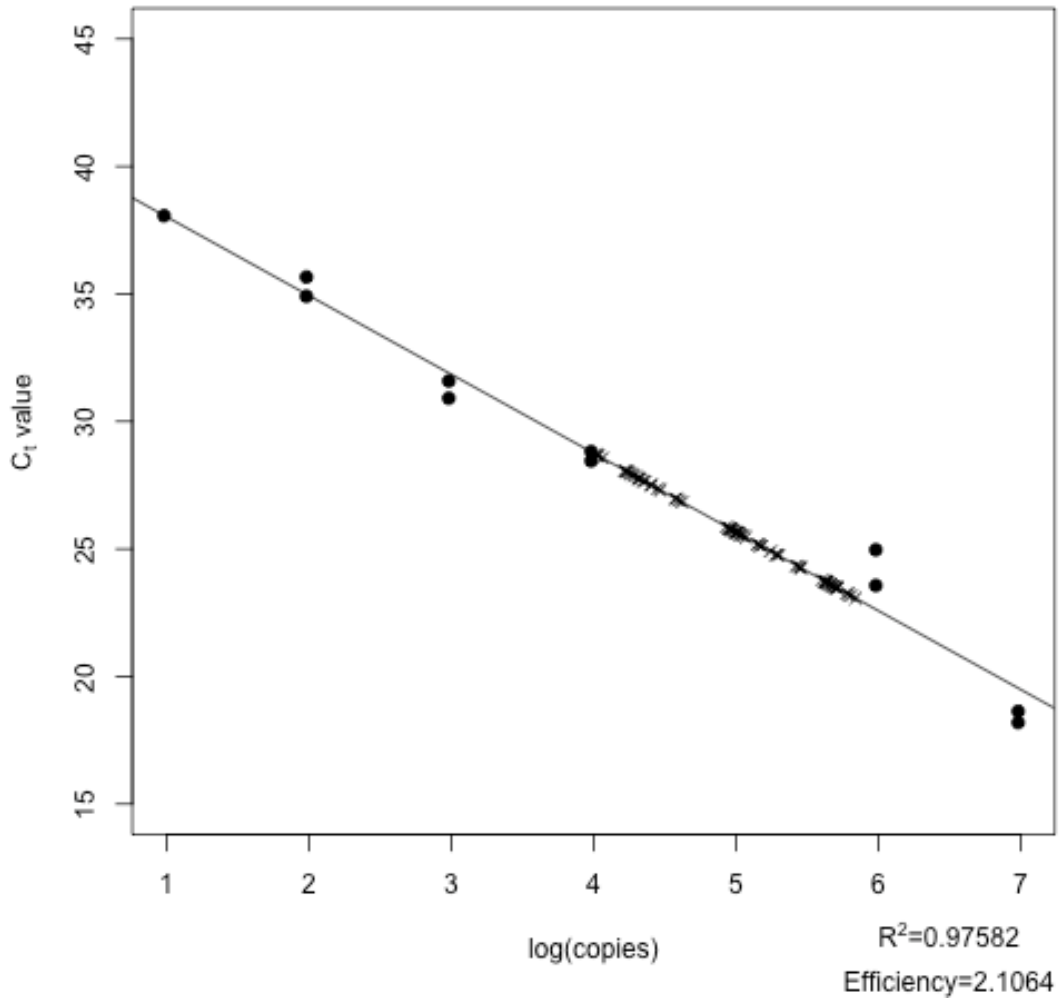


Figure 3.6 – Standard curve plot

A standard curve was plotted for an example plate. C_t values were plotted against the log-transformed copy number for seven different 1:10 dilutions of cloned control DNA (•). The C_t values for the unknown samples (×) were then overlaid on the standard curve and initial copy-number determined. Copy number was calculated using the linear regression equation.

$$correlation\ (X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Figure 3.7 – Pearson sample correlation coefficient

Given two genes X and Y measured with N conditions, the correlation between the two genes can be calculated using the above equation.

Results

After an initial analysis of the microarray data, 34 candidate genes were selected to be profiled using RT-PCR (Table 3.3). One of the candidates (Xl.13885.1.A1_s_at) could not be amplified with any cDNA template and was dropped. The remaining 33 candidate genes were profiled using semi-quantitative RT-PCR and visualized using agarose gel electrophoresis (Figure 3.8). Genes were ordered by their annotation classification.

Five of these genes were then selected for qPCR profiling by visual inspection of the RT-PCR profiles: Bix2, CHN1, Mix1, HBox1, and Zic5. Two of the selected genes were down-regulated with Nkx2-5 over-expression (Bix2 and Mix1), and the remaining three were up-regulated. Two were annotated as heart-related (Bix2 and Mix1), one was annotated as transcription-related (HBox1), and the last two were annotated as predominantly expressed in the head-region (CHN1 and Zic5). HBox1 had an expression pattern that seemed to mirror that of Nkx2-5, as did Zic5 (Figure 3.8). On the other hand, CHN1 was only found in heart cDNA. The rough expression pattern of Bix2 seemed to be the opposite of Nkx2-5, whereas Mix1 had an erratic expression pattern. Overall, it was felt that these genes represented a good cross-section of the original candidates.

Additionally, Nkx2-5 was included as a positive control and so that correlations to its expression pattern could be made. ODC was included as a housekeeping reference gene and GAPDH was included as a negative control. GAPDH is commonly used as a housekeeping reference gene, but under the variety of developmental stages profiled, its expression is highly variable (Sindelka et al. 2006). The size and concentration of cloned

Table 3.3 – Genes selected for RT-PCR profiling

Probe set	Xl UniGene ID	Xl UniGene Name	Absent Present Filter	Fold change	Welch's t-test (Log2) p-value	Storey FDR (Log2 t-test)	nkx2-5 site count	Tags
Xl.1093.1.S1_at (*)	Xl.1093	nkx2.5-prov	1.000	1.784	0.008	0.077	7	go:heart upreg-tf inhead
Xl.397.1.S1_at (*)	Xl.397	bix2-A	1.000	-1.928	0.015	0.096	3	go:heart
Xl.824.1.S1_at (*)	Xl.824	Mix.1	1.000	-2.359	0.003	0.064	1	go:heart
Xl.866.1.S2_at	Xl.866	LOC398047	1.000	2.773	0.014	0.094	4	go:heart
AFFX-Xl-a1Act-3_s_at	Xl.1115	acta1-prov	0.833	7.785	0.008	0.077	6	go:heart
AFFX-Xl-bAct-5_s_at	Xl.4138	MGC52661	1.000	1.511	0.015	0.095	6	go:heart
Xl.8190.1.S1_at	Xl.8190	mix.2-A	1.000	-2.453	0.001	0.043	1	go:heart
Xl.13437.1.A1_at	Xl.13437	LOC398263	1.000	-1.668	0.002	0.056	3	go:heart
Xl.25598.1.S1_at	Xl.25598	LOC495124	1.000	-1.896	0.002	0.058	8	go:heart
Xl.15970.1.A1_at	Xl.12327	MGC80468	1.000	1.523	0.010	0.083	6	go:heart upreg-tf
Xl.146.1.S1_at	Xl.146	XMyf-5	1.000	2.160	0.001	0.052	6	upreg-tf
Xl.481.1.S1_at	Xl.481	Gli2	0.917	1.818	0.010	0.082	1	upreg-tf

Table 3.3 (continued)

Probe set	Xl UniGene ID	Xl UniGene Name	Absent Present Filter	Fold change	Welch's t-test (Log2) p-value	Storey FDR (Log2 t-test)	nkx2-5 site count	Tags
Xl.580.1.S1_at	Xl.580	masking	1.000	1.650	0.005	0.071	2	upreg-tf
Xl.823.1.S1_at	Xl.823	LOC397758	1.000	1.911	0.006	0.072	6	upreg-tf
Xl.933.1.S1_at	Xl.933	bra3-A-prov	1.000	1.833	0.005	0.070	11	upreg-tf
Xl.975.1.S1_at	Xl.975	Tbx3	0.917	2.049	0.004	0.064	6	upreg-tf
Xl.1209.1.S1_at (*)	Xl.1209	XlHbox1	1.000	2.408	0.005	0.068	8	upreg-tf
Xl.1394.1.S1_at	Xl.1394		1.000	1.661	0.006	0.072	4	upreg-tf
Xl.16644.1.S1_at	Xl.16644	MGC84001	1.000	2.519	0.001	0.053	6	upreg-tf
Xl.18750.1.S1_at	Xl.18750	MGC53355	0.917	2.036	0.001	0.039	5	upreg-tf
Xl.19790.1.S1_at	Xl.19790	MGC82544	1.000	1.831	0.004	0.064	7	upreg-tf
Xl.20765.1.A1_at	Xl.20765	MGC80929	1.000	1.942	0.001	0.053	8	upreg-tf
Xl.21901.1.S1_at (§)	Xl.21901	ankrd3-prov	0.833	1.654	0.009	0.080	8	upreg-tf
Xl.25289.1.A1_at	Xl.50380	HSF2	1.000	1.521	0.003	0.064	5	upreg-tf
Xl.793.1.A1_at	Xl.793	XGATA-3	1.000	2.128	0.001	0.046	4	upreg-tf inhead

Table 3.3 (continued)

Probe set	Xl UniGene ID	Xl UniGene Name	Absent Present Filter	Fold change	Welch's t-test (Log2) p-value	Storey FDR (Log2 t-test)	nkx2-5 site count	Tags
Xl.1043.1.S1_at (*)	Xl.1043	zic5-A	1.000	1.706	0.015	0.095	9	inhead
Xl.1685.1.S1_at	Xl.1685	LOC398260	1.000	2.602	0.006	0.071	7	inhead
Xl.6393.1.S1_at	Xl.6393	LOC398081	1.000	-1.968	0.009	0.079	9	inhead
Xl.8333.1.A1_at	Xl.8333		1.000	-2.075	0.000	0.031	1	inhead
Xl.9113.1.A1_at (*)	Xl.9113	chn1-prov	1.000	2.098	0.001	0.043	5	inhead
Xl.12160.1.S1_at	Xl.12160	tnrc4-prov	1.000	2.301	0.008	0.078	4	inhead
Xl.15793.1.A1_at	Xl.15793		1.000	1.897	0.007	0.075	6	inhead
Xl.16169.1.S1_at	Xl.16169		1.000	1.780	0.007	0.072	3	inhead
Xl.13885.1.A1_s_at	Xl.29224		1.000	1.680	0.011	0.085	8	inhead

(*) Genes selected for qPCR profiling

(§) Gene dropped due to lack of PCR product

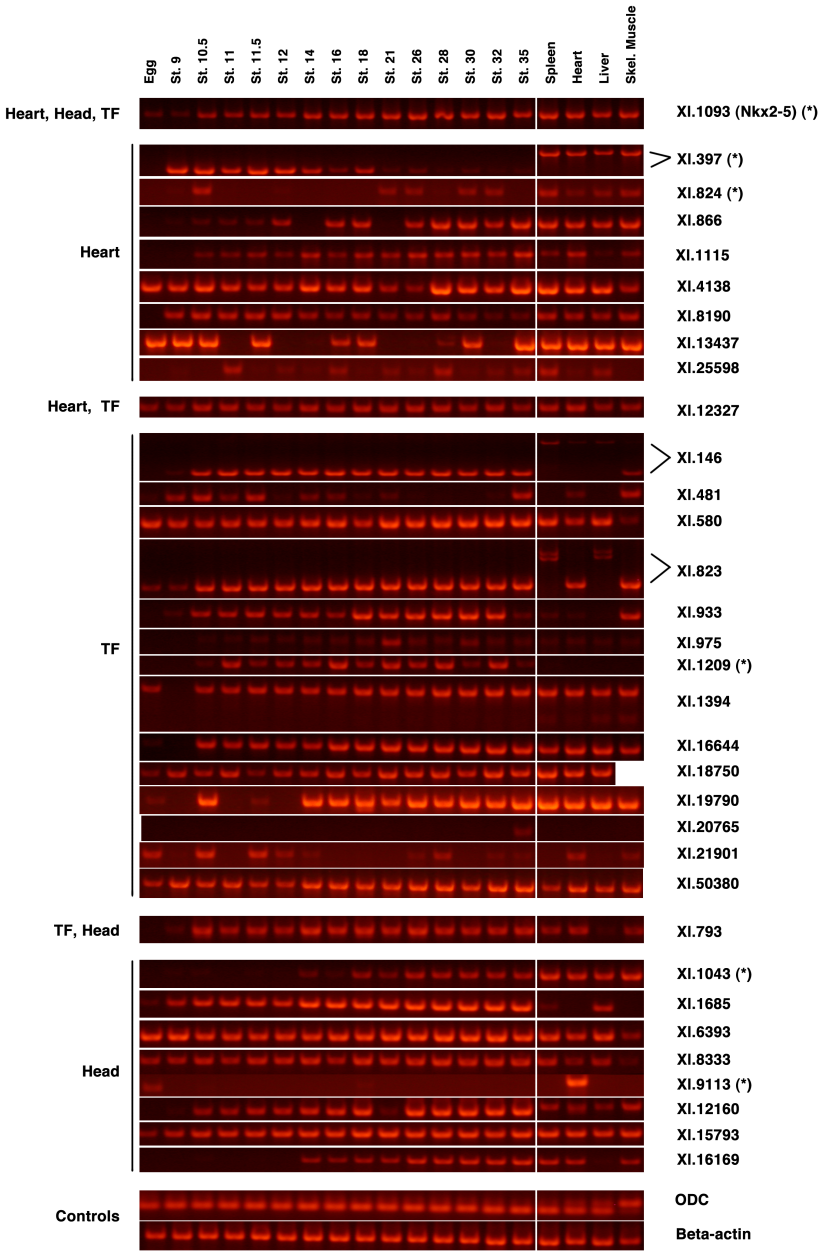


Figure 3.8 – Semi-quantitative RT-PCR profile of selected genes

Selected genes were profiled using semi-quantitative RT-PCR with cDNA from a variety of developmental stages and adult tissues. Genes are identified using their UniGene cluster IDs and marked on the left side by their classification: heart-related (Heart), predominantly expressed in the head (Head), or transcription-related (TF). The housekeeping genes ODC and β-actin were used as controls. Genes selected for qPCR profiling marked with (*). Three genes unexpectedly had multiple bands present (Xl.397, Xl.146, and Xl.823). It is unknown if these bands represented different splice variants or contamination. For all samples, 35 cycles of PCR amplification were performed.

standards were used to calculate the number of copies present in the undiluted standards (Table 3.4).

The selected 8 genes were then qPCR profiled using a serial dilution standard curve and cDNA from stages: St. 11, St. 17, St. 22, St. 28, and St. 45. Standard curves were calculated using linear regression (Figure 3.9) and the number of copies present in each of the staged cDNAs was calculated (Table 3.5). These values were normalized to the amount of ODC present in each biological replicate (Figure 3.10). The expression patterns of each gene were then correlated to the expression pattern of Nkx2-5 (Figure 3.11, Table 3.6).

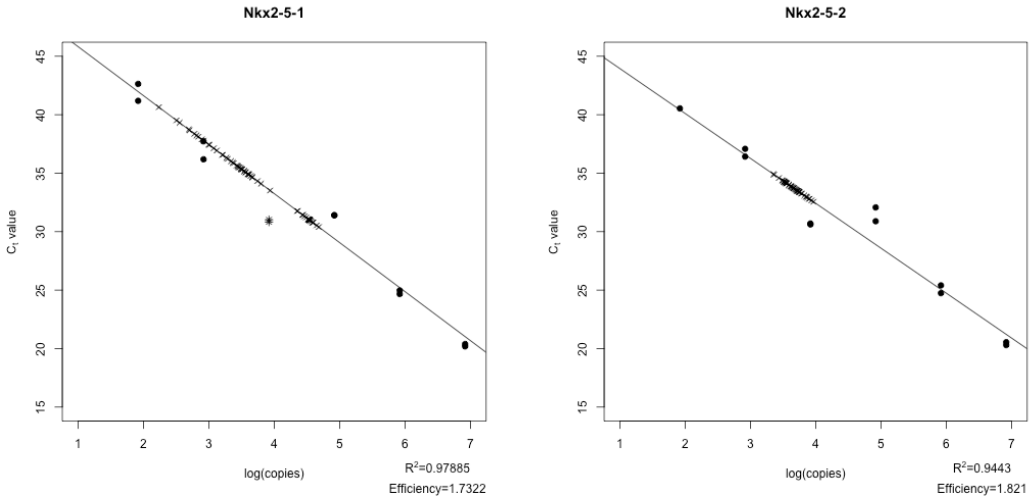
The expression patterns of Zic5 and HBox1 both correlated well to the pattern of Nkx2-5, whereas Mix1 and Bix2 were both down-regulated and not correlated. CHN1 was only marginally correlated, but still had a very large fold change, which might indicate that there was a secondary mechanism driving the changes in expression of CHN1.

Table 3.4 – The number of copies present in the control standard curves

The total sizes of the linearized control plasmids were calculated by adding the sizes of the insert with the size of the pCRII plasmid (3,973 bp). Then, using the equations in Figure 3.2, the measured concentrations, and a constant volume of 4 μl the number of copies in the undiluted control samples were calculated.

Gene	Insert (bp)	Total (bp)	Mass g/copy	Conc (ng/μl)	Copies
Nkx2-5	196	4,169	4.57E-018	94.8	8.299E+10
Bix2	182	4,155	4.55E-018	110.9	9.741E+10
CHN1	169	4,142	4.54E-018	99.4	8.758E+10
HBox1	209	4,182	4.58E-018	109.6	9.565E+10
Mix1	188	4,161	4.56E-018	107.4	9.420E+10
Zic5	147	4,120	4.52E-018	130.4	1.155E+11
ODC	221	4,194	4.60E-018	110.6	9.624E+10
GAPDH	195	4,168	4.57E-018	113.3	9.921E+10

Nkx2-5



Bix2

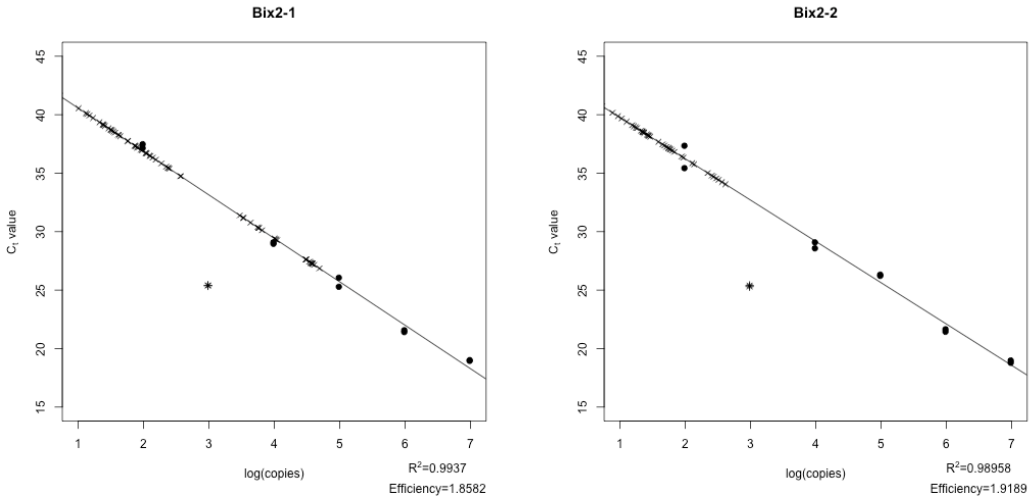
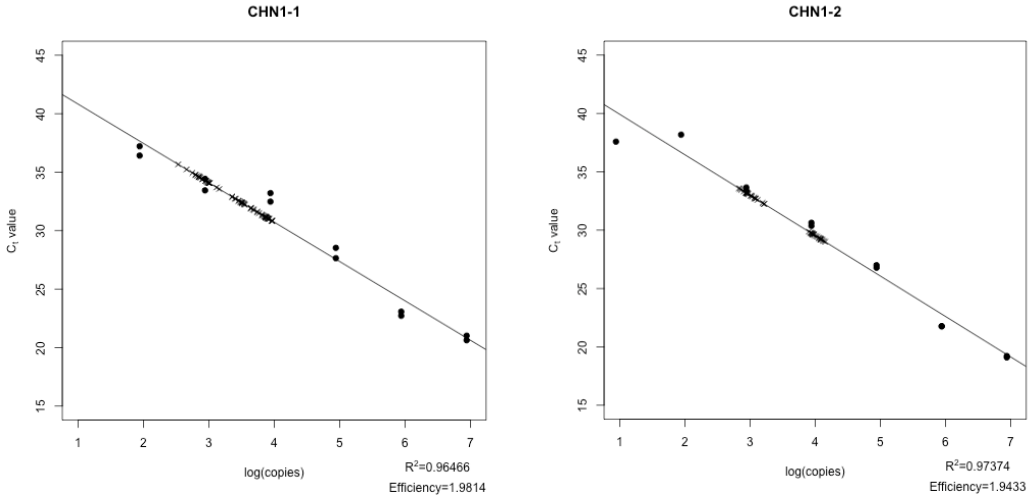


Figure 3.9 – Standard curves for qPCR profiled genes

Standard curves were plotted for each plate from each profiled gene. Each gene required two plates and standard curves were calculated for each plate independently. The cloned control plasmids were serial diluted (•) to form a standard curve. C_t values for the unknown samples were then projected onto the standard curve (×) and the number of copies in each of the staged cDNAs was calculated. Outliers in the serial dilutions were observed manually and disregarded (*).

CHN1



HBox1

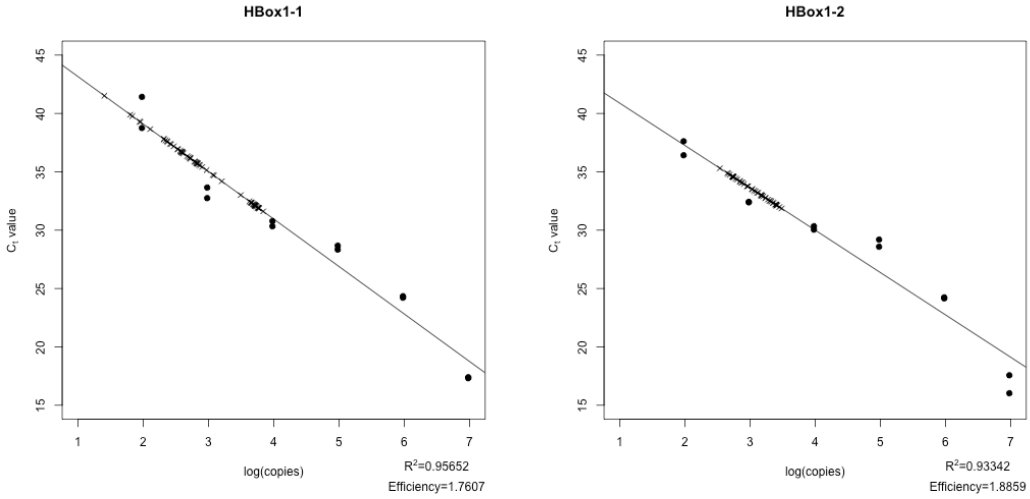
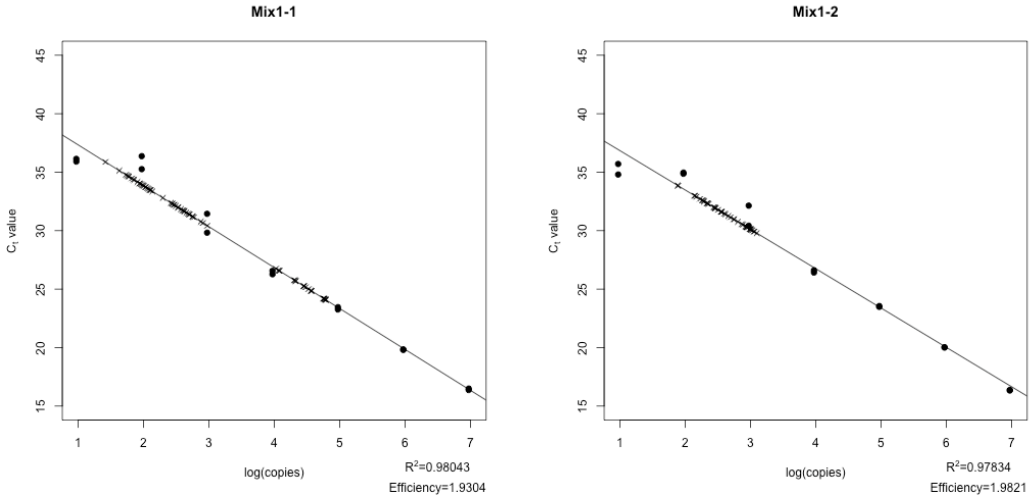


Figure 3.9 (continued)

Mix1



Zic5

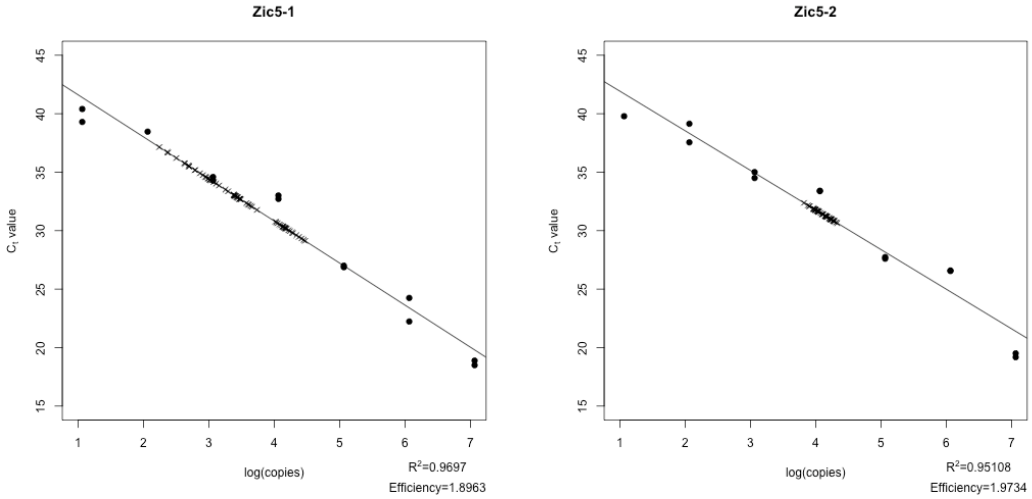
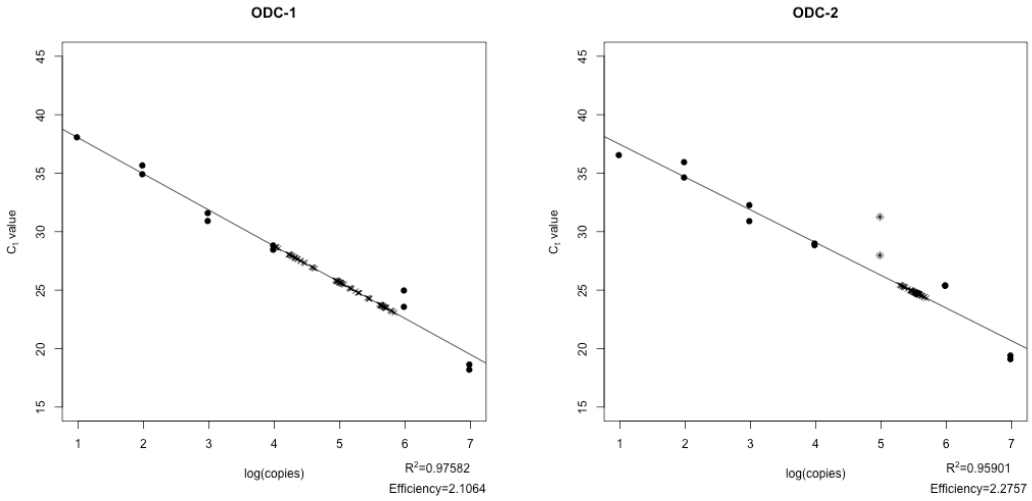


Figure 3.9 (continued)

ODC



GAPDH

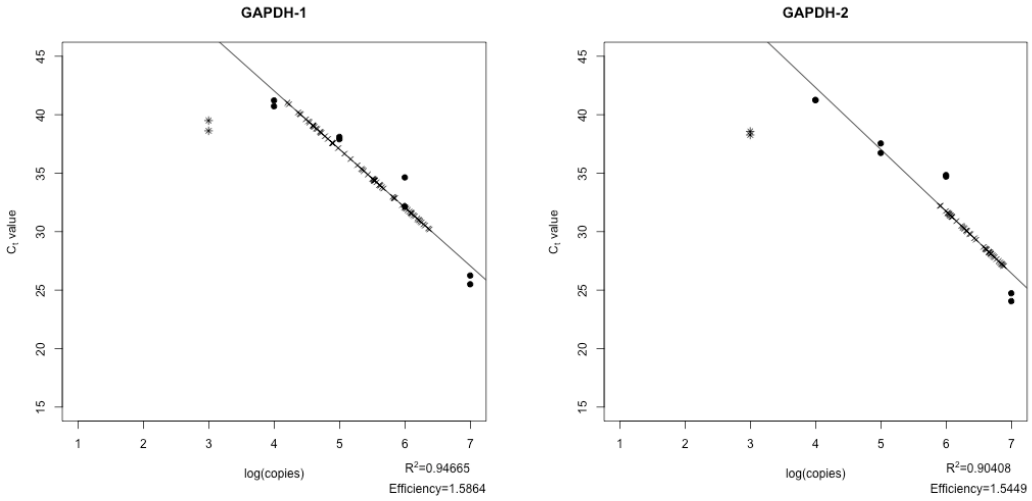


Figure 3.9 (continued)

Table 3.5 – Copy number for selected genes

	Nkx2-5	Bix2	CHN1	HBox1	Mix1	Zic5	ODC	GAPDH
St. 11-1	1315	10544	3284	178	36051	1028	146868	333016
St. 11-2	3716	5910	4954	1232	20818	1120	190471	253871
St. 11-3	492	41593	4993	248	28802	938	89421	470825
St. 11-4	520	31606	3211	59	58443	304	104467	384909
St. 11-5	555	37096	3143	68	59567	501	102698	338520
St. 11-6	2015	3517	2408	601	11549	606	101116	137151
St. 17-1	2736	274	1197	657	95	3723	39401	73781
St. 17-2	3827	21	782	299	85	2367	19114	40233
St. 17-3	3568	32	785	447	89	2528	17212	36900
St. 17-4	6205	43	872	641	90	4331	27145	55744
St. 17-5	2862	16	742	484	75	2694	21856	43983
St. 17-6	2126	50	600	448	97	2647	10843	20327
St. 22-1	37821	115	8508	5182	474	14122	637549	1928515
St. 22-2	32862	93	8143	4109	276	15848	425333	1385426
St. 22-3	35589	30	7251	5225	495	11131	481968	1033089

Table 3.5 (continued)

	Nkx2-5	Bix2	CHN1	HBox1	Mix1	Zic5	ODC	GAPDH
St. 22-4	37723	129	7706	5151	451	18369	494996	740990
St. 22-5	35284	31	7600	5332	675	14290	278022	1266290
St. 22-6	24929	188	7528	5972	327	26959	442497	1857354
St. 28-1	5212	19	1431	874	230	12866	323750	2255267
St. 28-2	3110	6	920	567	429	9344	287149	1148119
St. 28-3	3580	18	1053	479	199	11844	381047	1079463
St. 28-4	5273	26	809	1034	689	15827	203137	2165593
St. 28-5	4114	13	1238	573	204	9207	290517	1564120
St. 28-6	3912	18	805	635	325	8761	226301	970698
St. 45-1	3149	57	9602	1974	109	9930	344665	5484087
St. 45-2	5088	73	8894	1628	991	15367	362951	3978861
St. 45-3	5602	72	10135	1980	800	17351	370760	4619144
St. 45-4	5840	220	9303	2298	909	12619	333108	4104881
St. 45-5	8257	352	12423	1499	1143	18996	403345	6861075
St. 45-6	4830	29	12864	2740	339	18550	460640	6477777

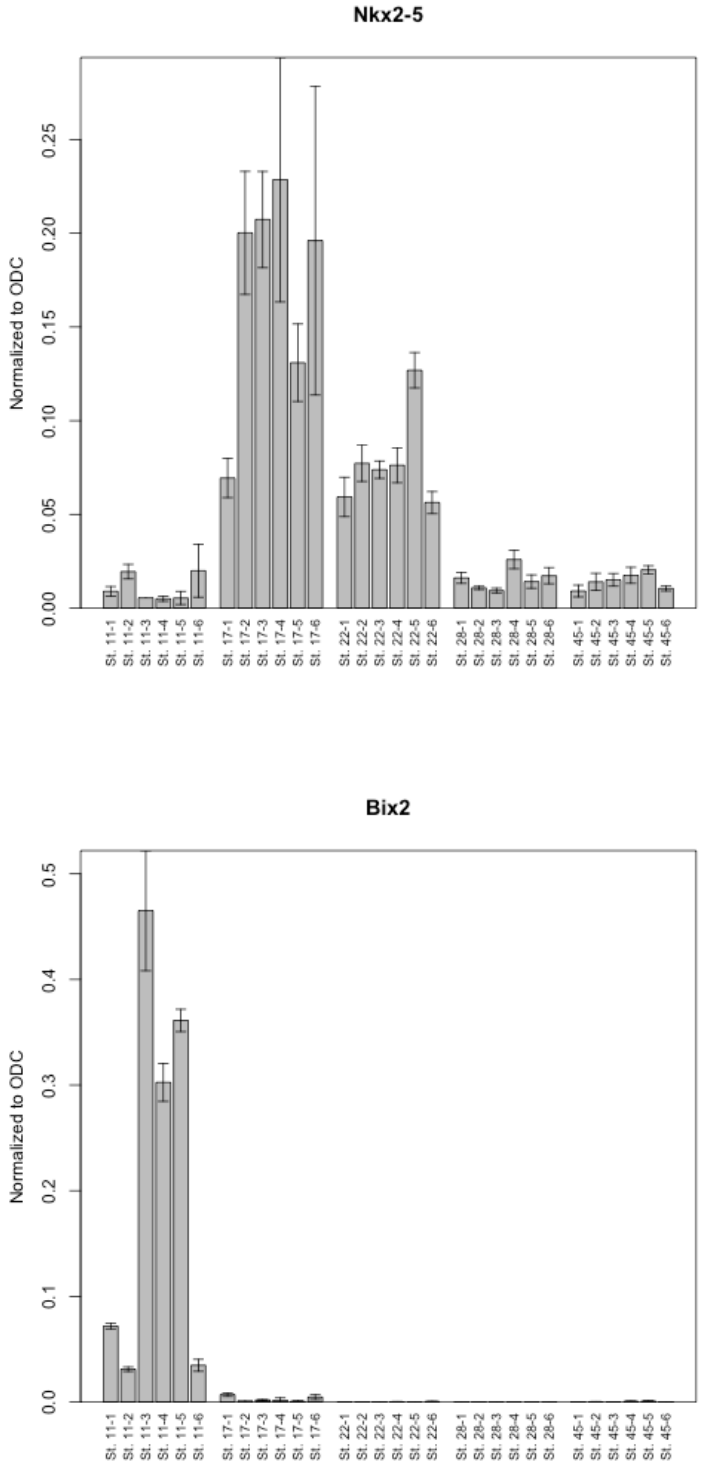


Figure 3.10 – qPCR expression profiles of selected genes (normalized to ODC)

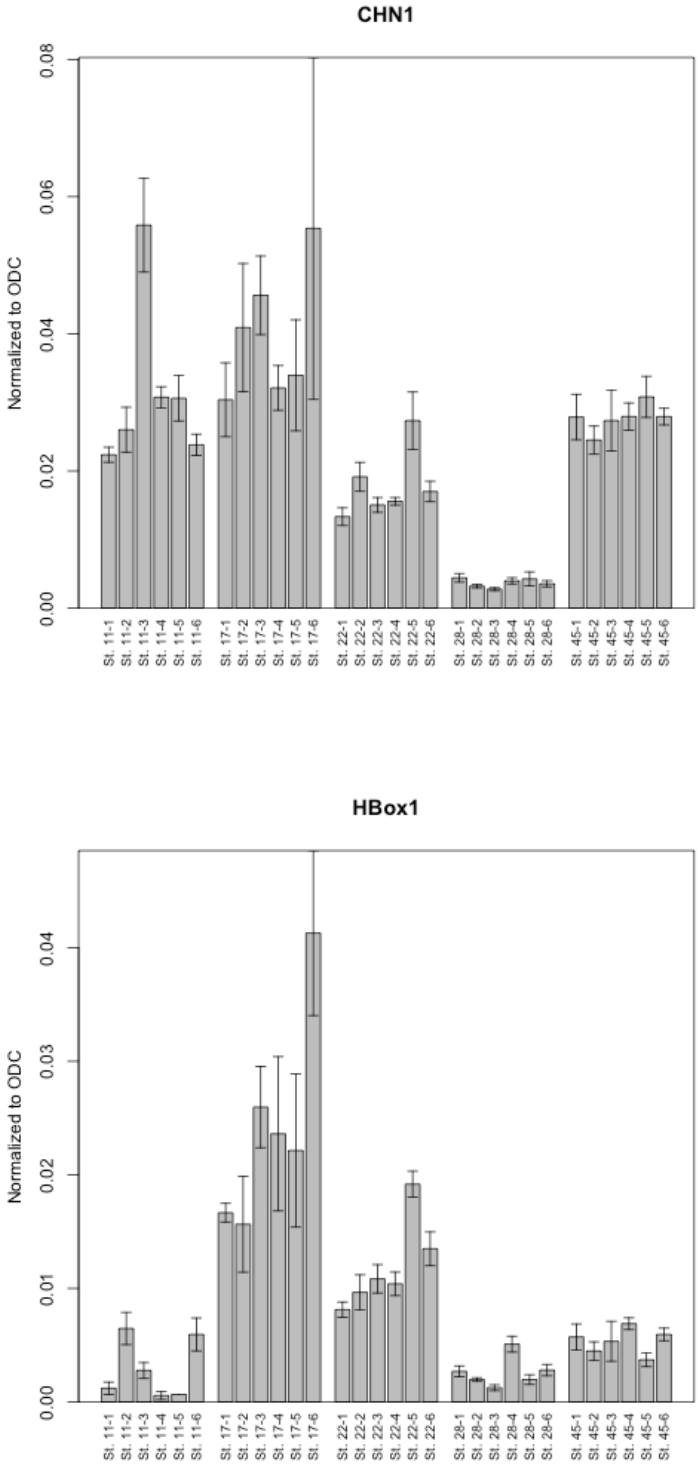


Figure 3.10 (continued)

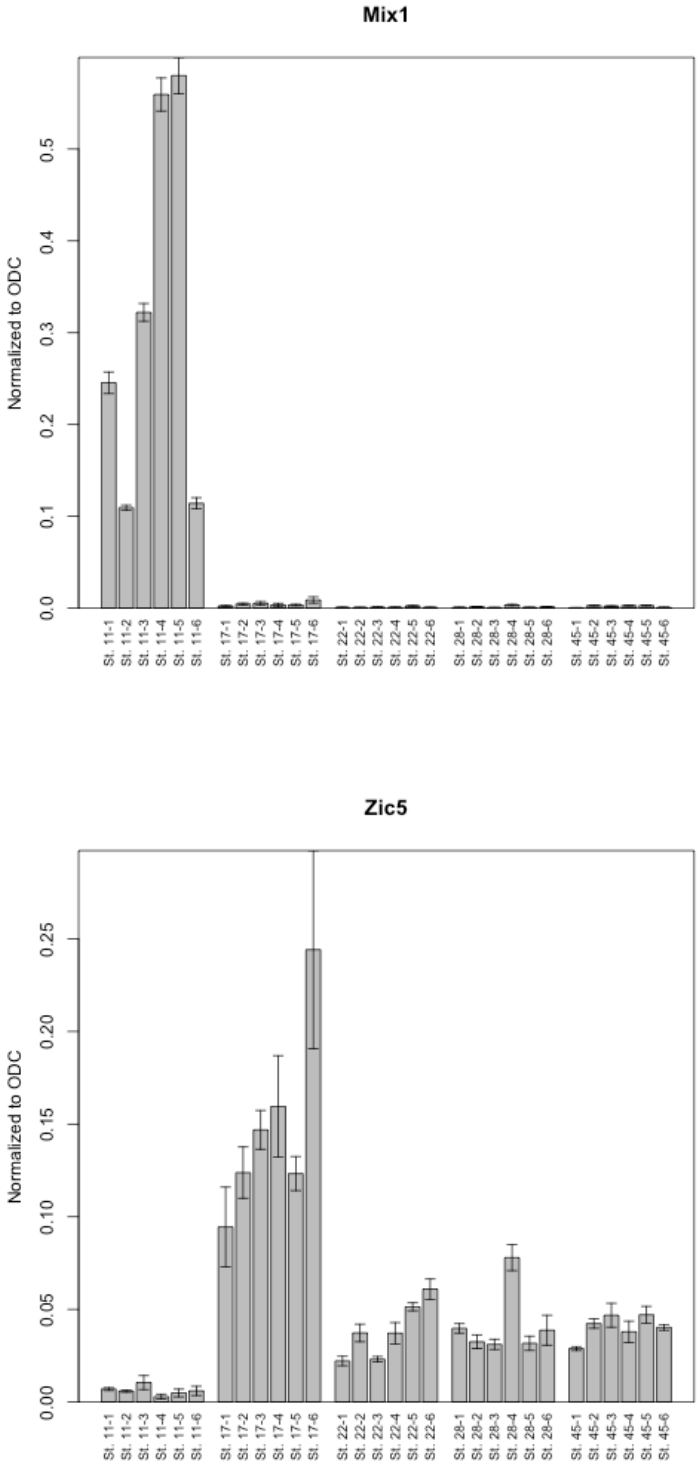


Figure 3.10 (continued)

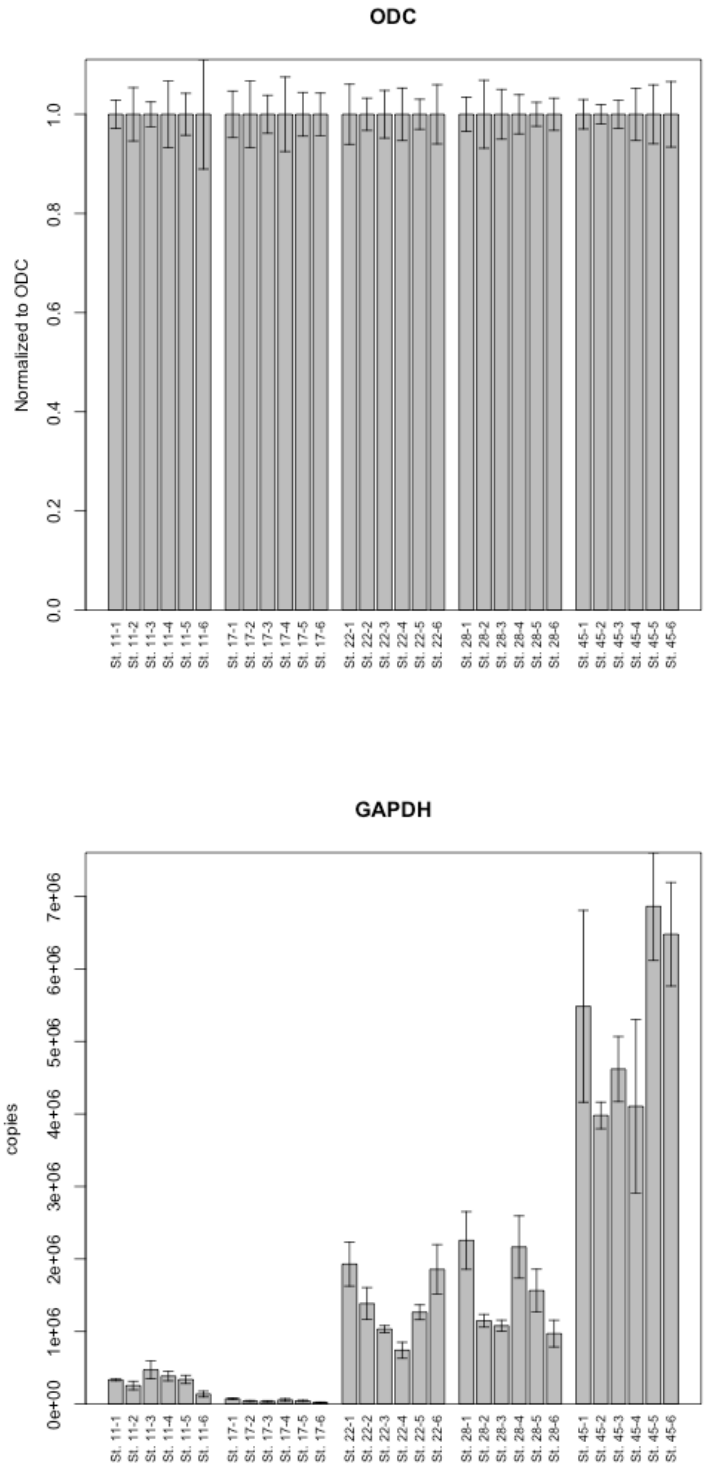


Figure 3.10 (continued)

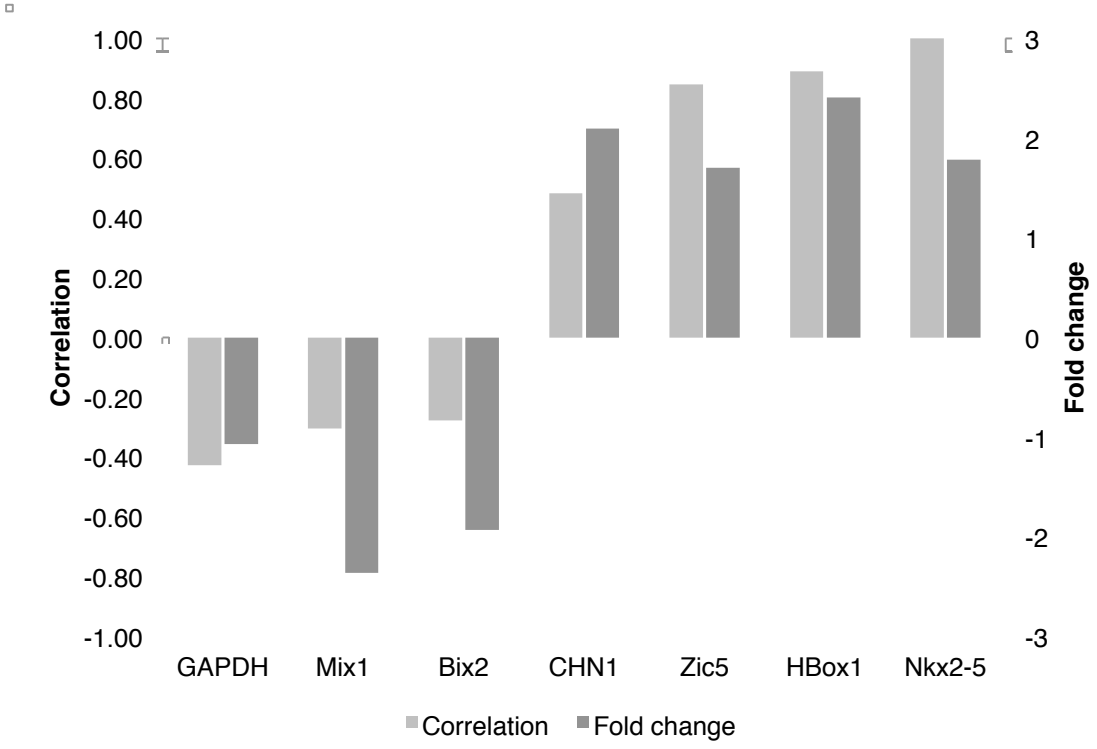


Figure 3.11 – Expression profile correlation with Nkx2-5

The Pearson correlation coefficient plotted for each of the profiled genes along with the fold change from the Nkx2-5 overexpression microarray (Chapter 2). The genes that had a greater fold change also had a better correlation to the expression profile of Nkx2-5.

Table 3.6 – Correlation of expression profiles to Nkx2-5

For each gene, the Pearson correlation coefficient was calculated, comparing its expression profile to that of Nkx2-5. Values greater than 0.6 are considered correlated.

Gene	Correlation
Bix2	-0.28
CHN1	0.48
HBox1	0.89
Mix1	-0.30
Zic5	0.85
GAPDH	-0.43

Discussion

Expression profiling is not a technique that can alone determine if genes are regulated by Nkx2-5. Profiling can imply that genes, which share a common expression pattern, may be regulated through similar mechanisms. Nkx2-5 is auto-regulatory, so Nkx2-5 may be the common regulator. But this guilt-by-association cannot be used to infer causality.

One thing that would have strengthened the profiling analysis would have been the ability to use a greater number of conditions. The semi-quantitative RT-PCR had 19 different stages and tissues. Because of this, the semi-quantitative profile gave better overview of the early expression of the 33 genes that were profiled. While the semi-quantitative profile gave a good overview, it wasn't accurate enough to be able to calculate expression pattern similarities.

The obvious choice then is to use more quantitative real-time PCR to quantify the exact number of copies of the target RNA molecule in each sample, as qPCR yields a much more accurate profile. To generate good data, it is important to use multiple biological replicates and multiple technical replicates. This was even more critical for low abundance genes where wells might be discarded when removing wells for quality control issues. Unfortunately, qPCR is an expensive technique to use to profile genes, limiting the number of genes and conditions available for testing.

The Roche LightCycler 480 has its own preferred technique for quantifying the amount of RNA in a sample: the second derivative max method (Wittwer et al. 1999; Luu-The et al. 2005). This method uses the maximum point of the second derivative in the growth of

fluorescence to determine a crossing point (C_p). The C_p is analogous to the C_t value that was calculated. The second derivative max method has two problems. First, it is very difficult to manually calculate, making it impossible to verify the results obtained from the software. Second, the method requires modeling the entire span of the exponential phase of the PCR reaction. Because of this, it may not be appropriate to use when measuring a target with very low abundance (on the order of 1,000 copies per sample). Because of these factors, it was determined that calculating C_t values was preferred. By calculating C_t values, it was possible to measure rare targets and still have good reproducibility. This was evident from the R^2 values of the standard curves. Every gene, except GAPDH, had a linear regression that fit the standard curve well. GAPDH measurements proved to be quite variable, and this reinforced why it was not chosen as a reference gene.

The qPCR analysis was limited by costs and reagent availability, but it nevertheless yielded some interesting results. The expression profiles of both Zic5 and HBox1 were highly correlated to the expression levels of Nkx2-5 and both were significantly up-regulated in the Nkx2-5 over-expression microarrays. Together, these two things imply that Nkx2-5 may have at least an indirect regulatory effect on these two genes.

CHAPTER 4: CONSTRUCTION AND USE OF THE CROSSGENE ANNOTATION DATABASE

Introduction

Modern studies in molecular biology increasingly require genomic scale analysis. The Achilles heel of genomic scale analysis is the quality of functional annotation, which varies widely from organism to organism. Specific organisms may be more widely studied in a particular domain, but not in others. The mechanisms of developmental biology, for instance, are often studied in model organisms such as fruit fly, mouse or frog and the relevant genes annotated primarily for these organisms. Genes associated with human disease will be annotated primarily in humans. This leads to gaps in annotation of even the most widely studied organisms.

One method for annotating unknown genes is to search for known protein domains within DNA or amino acid sequences. This approach is useful for a fraction of possible annotations, particularly cellular localization signals, protein-protein interaction domains, and DNA binding domains. It may miss larger biological processes or molecular functions. For example, finding a DNA binding domain suggests that a protein is likely to bind DNA, but does not reveal the effect of that binding. Additionally, orthologous proteins may gain or lose domains, and the effect of those domains may differ between organisms.

A common method for filling gaps in annotation is to compare gene or protein sequences from one organism to another. For example, to understand a gene expression experiment

carried out in the frog, *Xenopus laevis*, one might find the best match for each frog gene in human or mouse. However, this approach can also lead to problems. How does one best determine which organism to use as the reference? How does one set a threshold to say which genes are likely to be orthologs and which are questionable?

Another way to fill gaps in annotation is to find orthologous genes from multiple organisms. Examples include the Clusters of Orthologous Groups (COG) and euKaryotic clusters of Orthologous Groups (KOG) (Tatusov et al. 1997; Tatusov et al. 2003), the TIGR Orthologous Gene Alignments (TOGA) now known as Eukaryotic Gene Orthologues (EGO) (Lee et al. 2002), RoundUp (Deluca et al. 2006), OrthoMCL (Li et al. 2003), and HomoloGene (Wheeler et al. 2008). Each of these databases uses a different technique for finding orthologs. COG/KOG, RoundUp and OrthoMCL all use protein sequences as the basis of their comparisons. EGO is based upon tentative consensus (TC) sequences derived from ESTs. COG/KOG, EGO, and OrthoMCL all use a BLAST approach to find similar sequences, whereas RoundUp uses the reciprocal smallest distance algorithm (RSD) (Wall et al. 2003). Each of these has their strengths and weaknesses. COG/KOG and EGO haven't been updated since 2003 and 2006, respectively, and HomoloGene requires fully sequenced genomes. While RoundUp and OrthoMCL each have a large number of genomes (533 and 138 as of February 2010), the majority are prokaryotic and neither includes the common model organism, *Xenopus laevis*, in their index. Also, because of the large number of genomes, RoundUp orthologous clusters must be computed on the fly, which can be time consuming.

In an attempt to overcome the problems listed above and improve gene annotation, we developed the CrossGene database. CrossGene currently includes data from 8 commonly

studied eukaryotes: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. The full-length transcript sequences from each of these organisms were compared to all of the others using the BLAST family of algorithms (Altschul et al. 1990b; Altschul et al. 1997b). These homologies were then combined into one database. By comparing whole transcripts, we can identify orthologous genes or gene families across multiple species. Since genes with similar sequences are likely to have similar functions in different organisms, we then combine existing Gene Ontology (GO) annotations (Ashburner et al. 2000a; Consortium 2004a) for all members of the orthologous group to improve the functional annotation for each of the organisms included. Transferring annotation from one gene to another based on homology can have problems; for example, orthologs may differ in function between different organisms. However, relying on networks of reciprocal best-matches to establish cross-species homology, rather than on individual matches, should increase the reliability of the annotation.

Methods

Sequence retrieval and processing

UniGene clusters were treated as representative of the complete transcriptome for an organism (Wheeler et al. 2008). The unique UniGene cluster sequences were retrieved from NCBI (Table 4.1) and formatted into BLAST databases. Once in a common format, each sequence was compared to all of the databases. For cross-species

Table 4.1 – Sources of data included in CrossGene

Data set			Source	Date retrieved
H. sapiens	UniGene	218	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
M. musculus	UniGene	178	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
R. norvegicus	UniGene	179	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
G. gallus	UniGene	41	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
X. laevis	UniGene	86	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
D. rerio	UniGene	115	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
D. melanogaster	UniGene	63	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
C. elegans	UniGene	43	ftp://ftp.ncbi.nih.gov/repository/UniGene	2009-05-05
Gene Ontology Terms			http://genenontology.org/ontology	2009-06-09
UniProtKB identifier mappings			ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz	2009-06-09
EBI Gene Ontology Annotation			http://www.ebi.ac.uk/GOA	2009-06-09

(CrossGene build 0904)

comparisons, the TBLASTX algorithm was used. For comparisons within the same organism, the BLASTN algorithm was used. Raw blast results were parsed and stored in a MySQL relational database.

Best-match calculations

For each transcript, the blast hit with the lowest e-value for each other organism was defined as the "best-match." In the case of a tie, the blast hit with the best score was taken as the overall best-match. To further characterize the matches, if an e-value was within 10% of the log-transformed best-match e-value, it was deemed to be a "high-quality match." This was an arbitrary threshold. For this calculation, e-values of 0 were arbitrarily treated as 1e-200. For example, if the best-match of a transcript had an e-value of 1e-100, a high-quality match needed to have an e-value between 1e-100 and 1e-90. If two transcripts were best matches or high-quality matches for each other, they were called "reciprocal best-matches" or "reciprocal high-quality matches." These are also known as reciprocal best blast hits.

Reciprocal group assembly

Once best-matches and high-quality matches were calculated for each transcript, these results were used to calculate reciprocal groups for both best and high-quality matches. Any given transcript could belong to only one reciprocal best-match and one reciprocal high-quality group. For re-annotation, CrossGene assembled networks of reciprocal groups consisting linked reciprocal matches. For best-match groups, a new group was started with a random, unassigned, seed transcript. Then all of the reciprocal best matches for that transcript from other organisms were added to the group. Then all of the

reciprocal best matches for those transcripts were also added to the group. This continued recursively until all of the linked best-matches were added to the group. Because the graph was built by exhaustively traversing the reciprocal matches, the contents of a cluster were the same regardless of the order of seed transcripts. High-quality groups were formed in the same manner, except that they included linked high-quality reciprocal matches. Once a reciprocal group was assembled, the number of members from each organism was tallied. For reciprocal groups with more than one member but less than 100, a graphical representation was prepared using the dot and circo GraphViz layout programs (Gansner et al. 2000; Gansner et al. 2007).

GO annotation

The gene ontology (GO) term hierarchy was downloaded from the Gene Onotology Consortium (Ashburner et al. 2000a; Harris et al. 2004). GO terms were associated with transcripts by first mapping UniProtKB IDs to the UniGene cluster ID or alternatively the NCBI Gene ID of a CrossGene transcript using the UniProtKB "idmapping" datafile (Table 4.1). It was possible for multiple UniProtKB IDs to map to a single CrossGene transcript. Next, unfiltered GO annotations were retrieved from the EBI Gene Ontology Annotation database (GOA) (Camon et al. 2004a) for all organisms (Table 4.1). The UniProtKB identifier was then used to associate GO annotations to a CrossGene transcript. Because GO terms are in a hierarchy, reported annotation counts include non-redundant parent terms in their totals.

GO rescue and HomoloGene comparisons

For each organism, known GO terms were used to determine if they could be recovered using CrossGene reciprocal group GO annotations. Any GO annotation associated with a gene that was a member of non-singleton group was considered possible for recovery. If that annotation was present in another member of the reciprocal group (in a different organism), the annotation was rescued. Annotation evidence was also taken into account. If an annotation had a non-IEA evidence code and could be rescued by another non-IEA annotation, it was added to the non-IEA rescue counts.

Pairs of orthologous genes were extracted from NCBI HomoloGene and compared to CrossGene best-match and high-quality reciprocal groups. If both NCBI Gene IDs of a HomoloGene pair was present in CrossGene, it was considered a possible pair. If each gene in the pair were a member of the same best or high-quality group, the pair was confirmed. Possible and confirmed pairs were also categorized by organism pair.

Results

Interface and searching

CrossGene is available through a web interface with a section for individual transcripts and a section for reciprocal groups. The transcript section contains all of the previously known information for a transcript (Figure 4.1), including the sequence, other database identifiers (UniGene, GenBank, NCBI Gene, and UniProt), BLAST results (Figure 4.2), and existing GO term annotation. Each transcript also has a link to its best-match and high-quality reciprocal group. The interface for reciprocal groups is divided into three

areas: an overview (Figure 4.3), detailed member matches (Figure 4.4), and the GO term annotations for the group (Figure 4.5). Associated GO terms are also given a score, which is the number of evidence records linking the GO term to a member of the reciprocal group. The database can be queried with two methods: textual search or sequence based BLAST search. Any included database identifier, gene name, or keywords can be used to search the database for matches. The user can also upload a DNA sequence to compare with the included transcripts via the BLASTN algorithm.

In addition to the primary web interface, there is also a web API for retrieving data (Table 4.2). Complete annotations for all reciprocal groups and reciprocal group based GO annotations by organism can also be downloaded in a tab-delimited format.

Reciprocal group assembly

The number of members in a best-match group is highly variable (Table 4.3); however, each transcript belongs to one and only one group. Because high-quality groups were based on the less stringent reciprocal high-quality matches, they potentially yield much larger groups (Table 4.3) and graphs (Figure 4.6). While the majority of transcripts have no reciprocal best match, for each organism there is a core set of transcripts that have at least one reciprocal match (Table 4.4). In human, there are 19,862 and 23,406 transcripts with one or more best and high-quality reciprocal matches, respectively. These numbers correlate well with the estimated 20,000-25,000 protein coding genes present in the human genome (Consortium 2004b). Transcripts without a reciprocal match could indicate truly unique genes but more likely represent ESTs included in the UniGene database which are not protein coding.

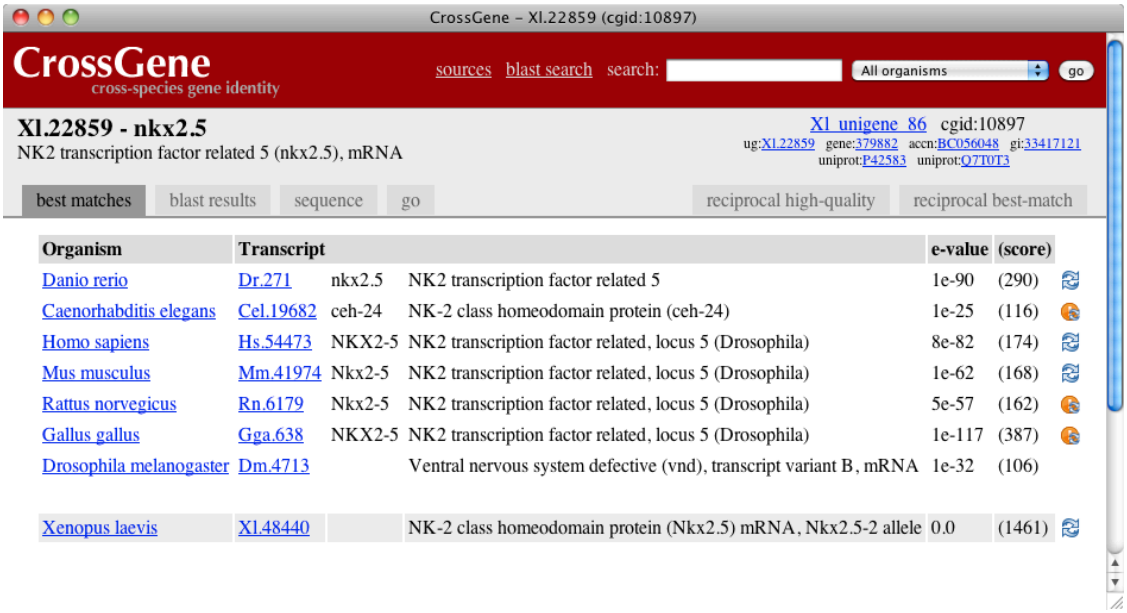


Figure 4.1 – Screenshot showing the best-matches and transcript overview

Listing of each of the best matches for the transcript Xl.22859 from all of the organisms in the CrossGene database. From this screen you can see the known information about the transcript, including the UniGene cluster ID, NCBI Gene ID, NCBI Entrez accession numbers, and UniProt IDs. The final column for each best-match indicates whether or not the match is a reciprocal best-match, reciprocal high-quality match, or non-reciprocal best or high-quality match.

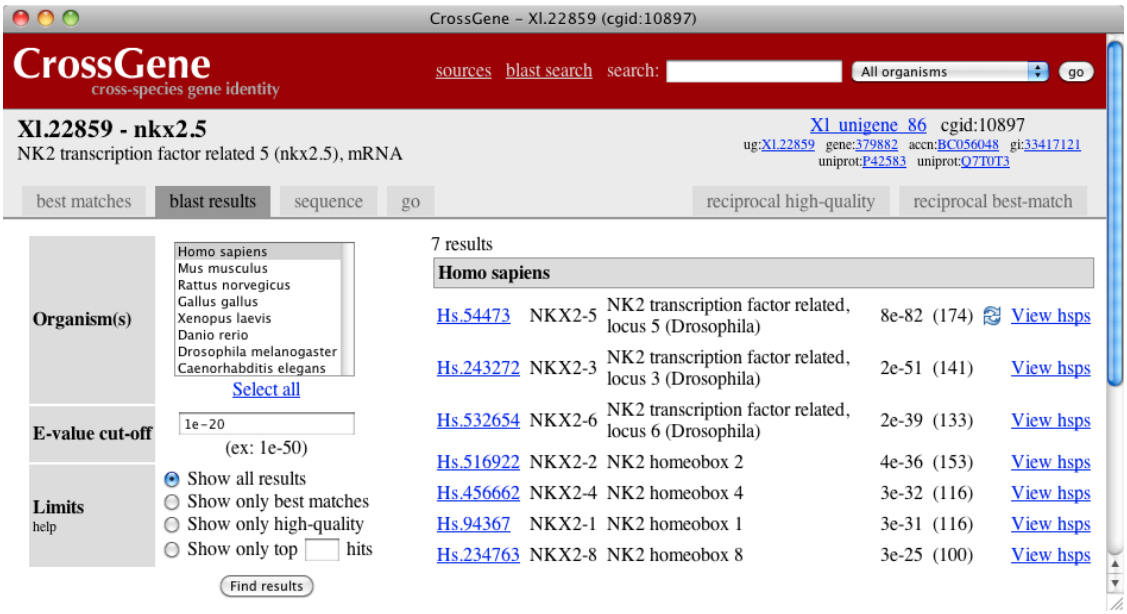


Figure 4.2 – Screenshot showing the BLAST results

BLAST results for Xl.22859 against *Homo sapiens*. Raw blast results can be filtered by one or more organisms, e-value, best/high quality match status, or limited to a defined number of hits per organism.

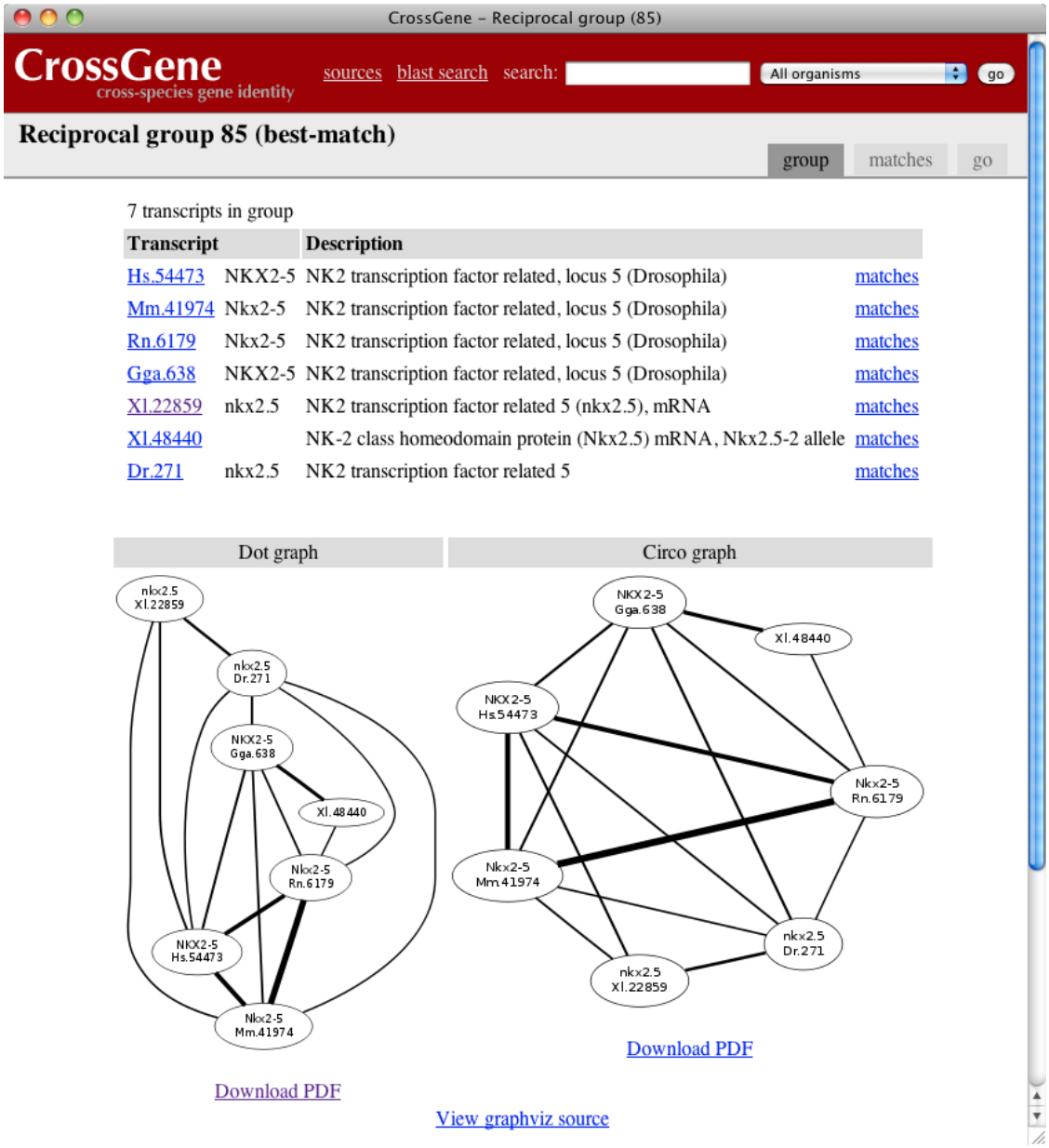
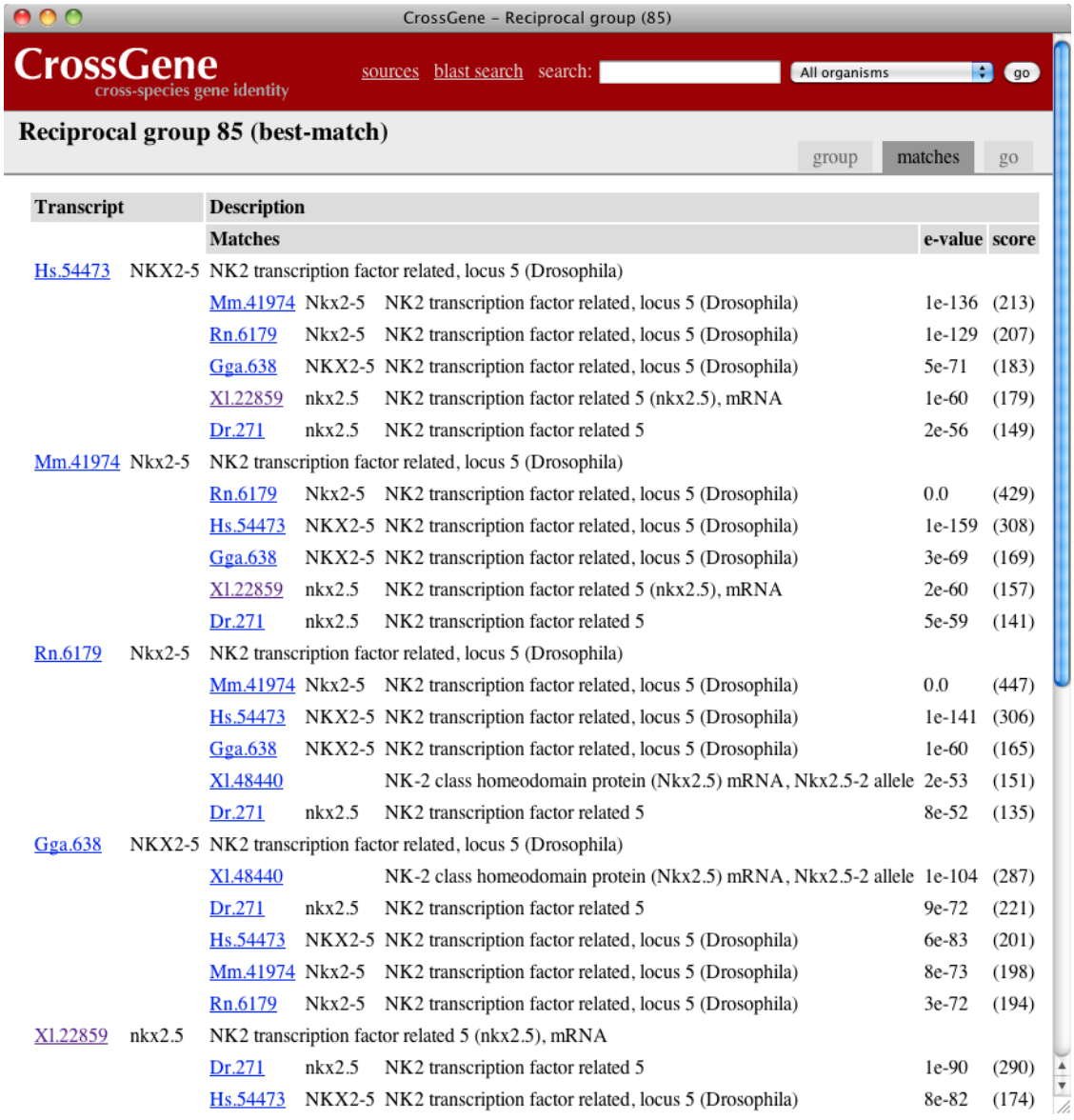


Figure 4.3 – Reciprocal group overview screen

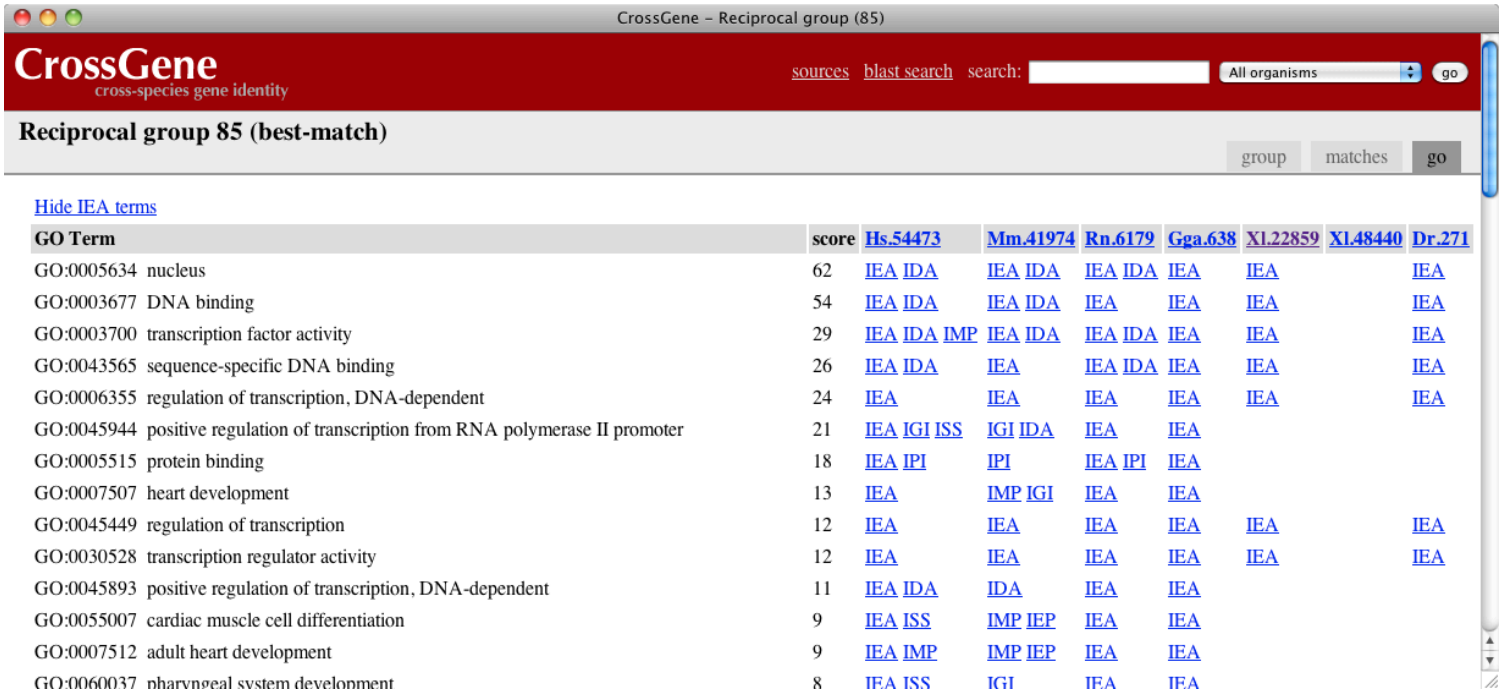
This is a screenshot of the reciprocal group overview screen for reciprocal group 85. This shows the members of reciprocal group 85 as well as the dot and circo graphs illustrating the network structure for this reciprocal group. Dot and Circo are two graph layout algorithms provided by the GraphViz package (Gansner et al. 2000; Gansner et al. 2007). Depending on the network structure, one may provide a more informative layout.



Transcript		Description				
		Matches			e-value	score
Hs.54473	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)				
		Mm.41974	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	1e-136	(213)
		Rn.6179	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	1e-129	(207)
		Gga.638	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	5e-71	(183)
		Xl.22859	nkx2.5	NK2 transcription factor related 5 (nkx2.5), mRNA	1e-60	(179)
		Dr.271	nkx2.5	NK2 transcription factor related 5	2e-56	(149)
Mm.41974	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)				
		Rn.6179	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	0.0	(429)
		Hs.54473	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	1e-159	(308)
		Gga.638	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	3e-69	(169)
		Xl.22859	nkx2.5	NK2 transcription factor related 5 (nkx2.5), mRNA	2e-60	(157)
		Dr.271	nkx2.5	NK2 transcription factor related 5	5e-59	(141)
Rn.6179	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)				
		Mm.41974	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	0.0	(447)
		Hs.54473	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	1e-141	(306)
		Gga.638	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	1e-60	(165)
		Xl.48440		NK-2 class homeodomain protein (Nkx2.5) mRNA, Nkx2.5-2 allele	2e-53	(151)
		Dr.271	nkx2.5	NK2 transcription factor related 5	8e-52	(135)
Gga.638	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)				
		Xl.48440		NK-2 class homeodomain protein (Nkx2.5) mRNA, Nkx2.5-2 allele	1e-104	(287)
		Dr.271	nkx2.5	NK2 transcription factor related 5	9e-72	(221)
		Hs.54473	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	6e-83	(201)
		Mm.41974	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	8e-73	(198)
		Rn.6179	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	3e-72	(194)
Xl.22859	nkx2.5	NK2 transcription factor related 5 (nkx2.5), mRNA				
		Dr.271	nkx2.5	NK2 transcription factor related 5	1e-90	(290)
		Hs.54473	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	8e-82	(174)

Figure 4.4 – Reciprocal group matches screen

This is a screenshot shows the matches screen for reciprocal group 85. This is an unordered list of reciprocal matches that were traversed to build this reciprocal group.



GO Term	score	Hs.54473	Mm.41974	Rn.6179	Gga.638	Xl.22859	Xl.48440	Dr.271
GO:0005634 nucleus	62	IEA IDA	IEA IDA	IEA IDA	IEA	IEA		IEA
GO:0003677 DNA binding	54	IEA IDA	IEA IDA	IEA	IEA	IEA		IEA
GO:0003700 transcription factor activity	29	IEA IDA IMP	IEA IDA	IEA IDA	IEA	IEA		IEA
GO:0043565 sequence-specific DNA binding	26	IEA IDA	IEA	IEA IDA	IEA	IEA		IEA
GO:0006355 regulation of transcription, DNA-dependent	24	IEA	IEA	IEA	IEA	IEA		IEA
GO:0045944 positive regulation of transcription from RNA polymerase II promoter	21	IEA IGI ISS	IGI IDA	IEA	IEA			
GO:0005515 protein binding	18	IEA IPI	IPI	IEA IPI	IEA			
GO:0007507 heart development	13	IEA	IMP IGI	IEA	IEA			
GO:0045449 regulation of transcription	12	IEA	IEA	IEA	IEA	IEA		IEA
GO:0030528 transcription regulator activity	12	IEA	IEA	IEA	IEA	IEA		IEA
GO:0045893 positive regulation of transcription, DNA-dependent	11	IEA IDA	IDA	IEA	IEA			
GO:0055007 cardiac muscle cell differentiation	9	IEA ISS	IMP IEP	IEA	IEA			
GO:0007512 adult heart development	9	IEA IMP	IMP IEP	IEA	IEA			
GO:0060037 pharyngeal system development	8	IEA ISS	IGI	IEA	IEA			

Figure 4.5 – GO annotations screen

This screenshot shows the GO term annotations for reciprocal group 85, including IEA annotations. GO terms are ordered by the number of times a GO term is associated to a transcript in this group (score). On the right of the screen is a column for each member of the group containing links to the evidence of this GO term annotation to the member.

Table 4.2 – HTTP API URLs

Data	URL
Reciprocal group members	http://crossgene.cmg.iupui.edu/recip/[id]?output=text
Reciprocal group GO annotations	http://crossgene.cmg.iupui.edu/recip/[id]/go?output=text
Transcript annotation	http://crossgene.cmg.iupui.edu/transcript/[id]?output=text
Transcript sequence	http://crossgene.cmg.iupui.edu/transcript/[id]/seq?output=text
Transcript GO annotation	http://crossgene.cmg.iupui.edu/transcript/[id]/go?output=text
List of all sources	http://crossgene.cmg.iupui.edu/sources?output=text
Transcripts for a source	http://crossgene.cmg.iupui.edu/source/[id]?output=text

Table 4.3 – Size of best-match and high-quality reciprocal groups

Group members	Best-match	High-quality
1000+	-	1 (1487)
400-1000	-	4
201-400	-	10
101-200	-	30
65-100	1 (97)	29
51-65	9	47
41-50	11	64
31-40	30	140
26-30	51	174
21-25	168	317
16-20	479	764
11-15	1,624	1,653
7-10	4,549	3,903
3-6	7,441	5,879
2	8,094	7,508
1	303,773	283,707

GO annotation

When looking at the total number of annotated transcripts, using best-match reciprocal groups to combine annotations yielded more annotated transcripts for all organisms (Table 4.5). As expected, the increase in number of annotated transcripts was greatest in organisms that originally had a lower percentage of annotated transcripts. For example, in *Caenorhabditis elegans* and *Drosophila melanogaster*, the two best-annotated organisms by percentage, there was only a net gain of 760 and 716 annotated transcripts, respectively. However, *Xenopus laevis* gained 4,260 additional annotated transcripts and *Danio rerio* gained 6,071. The number of annotated transcripts almost tripled for *Gallus gallus*. Even as well studied an organism as *Rattus norvegicus* gained 8,173 annotated transcripts, nearly doubling the previous annotations.

In addition to increasing the number of annotated transcripts, the average number of annotations for each annotated transcript also increased dramatically for all organisms. Increases ranged from 46% in humans to over 280% in *Xenopus laevis,* and on average they doubled (Table 4.5).

Robustness of GO annotations

The quality of the reciprocal group GO annotations was tested by exploring the ability of CrossGene to rescue the known annotations of each organism (Table 4.6, Table 4.7). When the annotations of an organism were completely excluded, how well could CrossGene reconstruct those annotations using the information from the other organisms? If a gene was a member of a reciprocal group of more than one member, any known

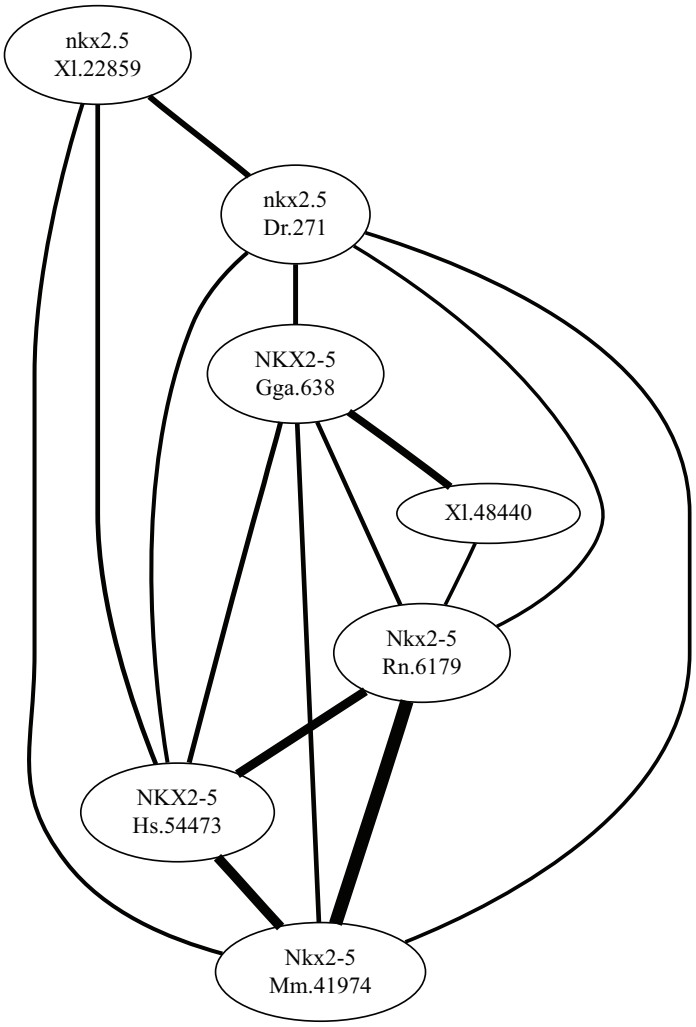


Figure 4.6 – Best-match reciprocal group for Nkx2-5

This is the graph representation of the best-match reciprocal groups containing the *Xenopus laevis* gene Nkx2-5. The widths of the edges are inversely proportional to the e-value of the match. The best-match graph shows all of the cross-species genes homologous to Nkx2-5.

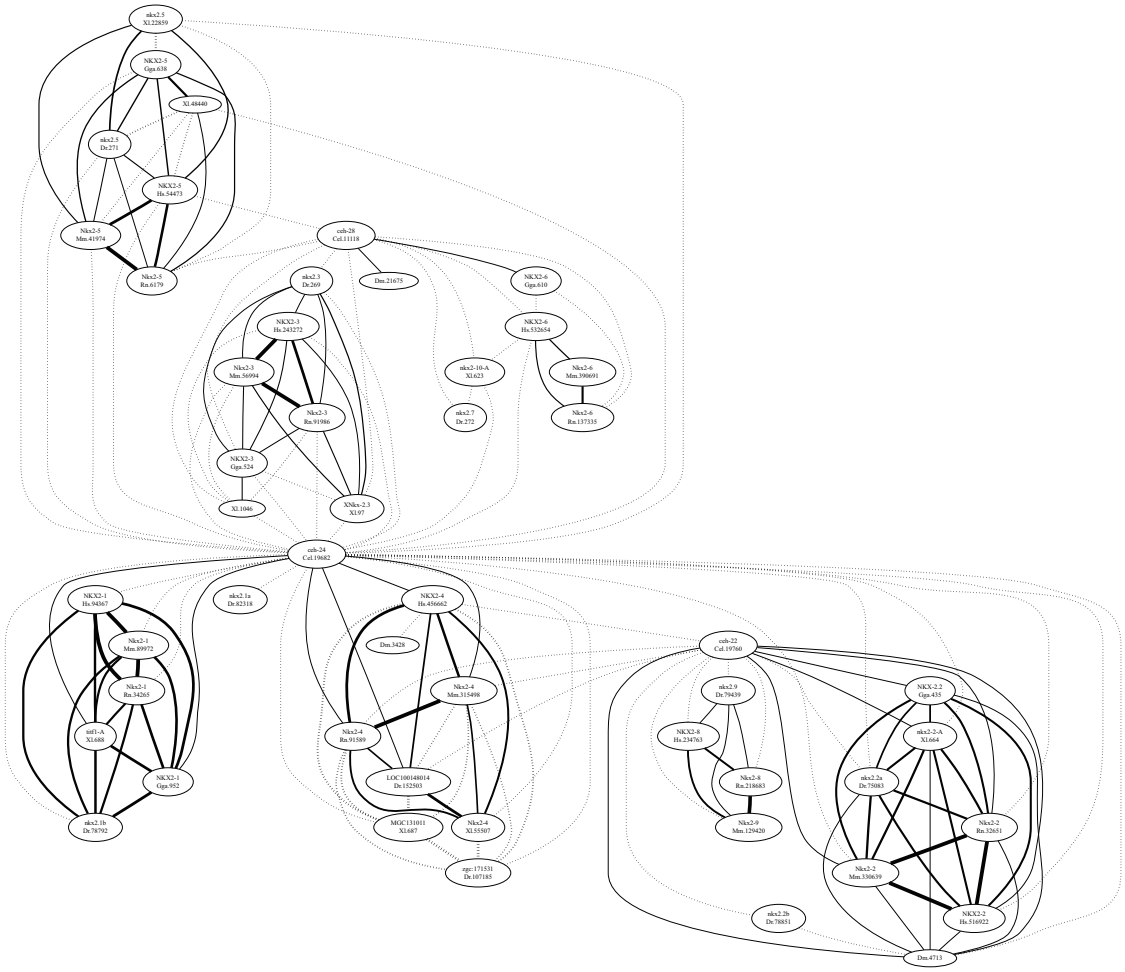


Figure 4.7 – High-quality reciprocal group for Nkx2-5

This is the graph representation of the high-quality reciprocal groups containing the *Xenopus laevis* gene Nkx2-5. The widths of the edges are inversely proportional to the e-value of the match. Best-matches are drawn in solid lines and high-quality matches are drawn in dashed lines. The best-match graph (Figure 4.6) shows all of the cross-species genes homologous to Nkx2-5, whereas the high-quality group has expanded the network to include other *Nkx2* family members in all other organisms, including Nkx2-1 through Nkx2-10.

Table 4.4 – Transcripts with at least one reciprocal best or high-quality match

Organism	# seq	# best-match	# hq-match
H. sapiens	123,877	19,862 (16.0%)	23,406 (18.9%)
M. musculus	79,021	23,852 (30.2%)	28,220 (35.7%)
R. norvegicus	63,440	23,605 (37.2%)	26,816 (42.3%)
G. gallus	33,383	13,999 (41.9%)	15,198 (45.5%)
X. laevis	35,077	13,111 (37.4%)	15,393 (43.9%)
D. rerio	51,506	15,464 (30.0%)	19,324 (37.5%)
D. melanogaster	17,330	6,375 (36.8%)	7,137 (41.2%)
C. elegans	22,015	5,608 (25.5%)	6,448 (29.3%)

Table 4.5 – Transcript annotation levels before and after CrossGene best-match reciprocal group annotation

Listed are the transcripts that have at least one GO annotation. Annotations are divided based upon their evidence codes: non-*inferred electronic annotation* (Non-IEA), only IEA, and all (Non-IEA and IEA). Finally, GO annotations derived from CrossGene best-match reciprocal groups are shown. Percentages are based upon the total number of transcripts included for that organism. Also shown is the average number of annotations for each annotated transcript. For all organisms, use of CrossGene best-match reciprocal groups enhanced the number of annotated transcripts and enhanced the breadth of annotation for each individual transcript.

		Number of annotated transcripts				Ave. annotations per transcript	
Organism	**# seq**	**Non-IEA**	**IEA**	**All**	**CG**	**All**	**CG**
H. sapiens	123,877	11,024 (8.9%)	6,310 (5.1%)	17,334 (14.0%)	19,919 (16.1%)	29.5	43.1
M. musculus	79,021	7,944 (10.1%)	9,603 (12.2%)	17,547 (22.2%)	20,834 (26.4%)	25.3	41.8
R. norvegicus	63,440	4,614 (7.3%)	5,347 (8.4%)	9,961 (15.7%)	18,134 (28.6%)	27.7	44.8
G. gallus	33,383	431 (1.3%)	3,910 (11.7%)	4,341 (13.0%)	12,914 (38.7%)	23.6	48.5
X. laevis	35,077	524 (1.5%)	8,894 (25.4%)	9,418 (26.8%)	13,678 (39.0%)	16.5	46.7
D. rerio	51,506	1,738 (3.4%)	9,791 (19.0%)	11,529 (22.4%)	17,600 (34.2%)	16.6	43.9
D. melanogaster	17,330	6,437 (37.1%)	3,342 (19.3%)	9,779 (56.4%)	10,495 (60.6%)	19.7	37.3
C. elegans	22,015	5,091 (23.1%)	6,101 (27.7%)	11,192 (50.8%)	11,952 (54.3%)	16.3	31.8

Table 4.6 – GO annotation rescue (best-match)

	All (incl. IEA)			Non-IEA		
	Rescued	**Possible**	**%**	**Rescued**	**Possible**	**%**
H. sapiens	88,847	151,481	58.7%	15,622	57,034	27.4%
M. musculus	87,423	125,867	69.5%	13,240	39,833	33.2%
R. norvegicus	61,712	78,096	79.0%	8,990	24,645	36.5%
G. gallus	25,211	28,067	89.8%	1,531	1,895	80.8%
X. laevis	26,827	34,881	76.9%	1,325	2,158	61.4%
D. rerio	29,164	46,950	62.1%	1,610	5,511	29.2%
D. melanogaster	24,640	58,666	42.0%	4,455	24,031	18.5%
C. elegans	18,568	51,425	36.1%	1,091	15,228	7.2%

Table 4.7 – GO annotation rescue (high-quality)

	All (incl. IEA)			Non-IEA		
	Rescued	**Possible**	**%**	**Rescued**	**Possible**	**%**
H. sapiens	78,113	129,746	60.2%	16,013	52,978	30.2%
M. musculus	76,188	103,665	73.5%	13,657	37,629	36.3%
R. norvegicus	54,404	67,314	80.8%	9,248	23,061	40.1%
G. gallus	23,055	24,988	92.3%	1,582	1,872	84.5%
X. laevis	24,336	27,326	89.1%	1,385	1,910	72.5%
D. rerio	27,063	36,478	74.2%	1,883	5,115	36.8%
D. melanogaster	24,394	54,517	44.7%	4,848	23,398	20.7%
C. elegans	18,159	47,659	38.1%	1,167	14,832	7.9%

annotations for that gene were considered as candidates for rescue. An annotation was rescued if it was present in any other members of the reciprocal group (in a different organism). Non-IEA annotations were tallied separately, because they may be based upon the excluded organism. To examine the performance using only non-IEA annotations, only non-IEA annotations were considered for rescue and were only rescued if annotated with non-IEA evidence in a different organism. Best-match (Table 4.6A) and high-quality reciprocal groups (Table 4.6B) were tested separately for robustness.

In general, each organism had significantly better recovery when all annotations could be used for recovery, and the less stringent high-quality reciprocal groups yielded more rescued annotations than best-match groups. The degree of recovery was different for each organism, and corresponds to the degree of unique annotations for that organism. For example, two of the organisms with the fewest non-IEA annotations, *Xenopus laevis* and *Gallus gallus*, had the best recovery using non-IEA annotations. However, other organisms faired less well in non-IEA recovery, indicating that there isn't yet significant overlap in non-automated annotation across these organisms.

HomoloGene ortholog comparison

It is also useful to compare the orthologs predicted by CrossGene reciprocal groups to those from other sources. One commonly used source is NCBI's HomoloGene database (Wheeler et al. 2008). HomoloGene merges genes into clusters based upon protein sequence similarity and genomic synteny, which requires sequenced genomes. Additionally, HomoloGene includes a mechanism for directly including paralogs into

their clusters[1]. CrossGene does not actively search for paralogs, although they can be included in the same reciprocal group if they are connected through other organisms. HomoloGene clusters (release 64) were retrieved from NCBI, with each gene identified by an NCBI GeneID. Not all organisms present in CrossGene were present in HomoloGene, or in the case of *Drosophila melanogaster*, CrossGene was lacking GeneID information. For each pair of genes in a cluster, CrossGene was searched for transcripts with the same GeneID. If both members of a pair were present in CrossGene, the pair was considered a possible match. If both transcripts belong to the same best or high-quality reciprocal group, the pair was confirmed by CrossGene. Confirmed and possible matches were classified by organism pairs and the percentage of confirmed matches was calculated (Tables 4.8-4.10).

Inter-organism matches compared well to HomoloGene, with between 61.2% and 85.2% of HomoloGene pairs corroborated by CrossGene best-match groups (Table 4.8). Using high-quality groups increased the matches to between 74.2% and 91.2%.

[1] http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html

Table 4.8 – HomoloGene confirmation percentage

	Best-match		High-quality	
	+ self	**- self**	**+ self**	**- self**
H. sapiens	77.8%	79.1%	85.0%	85.8%
M. musculus	56.4%	65.8%	70.2%	74.5%
R. norvegicus	74.1%	83.3%	83.7%	90.4%
G. gallus	79.5%	85.2%	86.1%	91.2%
D. rerio	12.0%	61.2%	19.2%	74.2%
C. elegans	11.4%	74.6%	15.5%	87.2%

Table 4.9 – Percentage of organism-to-organism pairs confirmed (best-match)

	Cel	Dr	Gga	Hs	Mm	Rn
Cel	-	71.8%	79.9%	71.5%	72.0%	79.2%
Dr	-	-	81.8%	72.7%	37.9%	76.5%
Gga	-	-	-	84.2%	84.8%	92.2%
Hs	-	-	-	-	78.0%	84.7%
Mm	-	-	-	-	-	83.1%

Table 4.10 – Percentage of organism-to-organism pairs confirmed (high-quality)

	Cel	Dr	Gga	Hs	Mm	Rn
Cel	-	91.0%	90.7%	82.3%	83.0%	89.5%
Dr	-	-	92.8%	84.6%	51.3%	88.9%
Gga	-	-	-	88.2%	88.7%	95.6%
Hs	-	-	-	-	83.4%	89.0%
Mm	-	-	-	-	-	90.0%

Discussion

Identification and annotation

The major goals of CrossGene are to identify orthologous genes between different species and to use those matches to expand the GO annotation of matched genes. Orthologs are determined by the reciprocal best-matches of genes between organisms. If two genes are reciprocal best-matches with each other, then it is reasonable to treat the two as orthologs, and by definition, orthologous genes are likely to have similar functions. Reciprocal high quality matches serve a similar purpose, and they allow the function of a gene to be predicted and the annotations enriched at a less stringent threshold.

Reciprocal best match GO annotations can be retrieved with and without inferred electronic annotations (IEA). IEA annotations are computationally produced, and have not been reviewed by a curator. Because CrossGene itself is a form of unsupervised computational annotation, the user may wish to exclude other IEA annotations in CrossGene reciprocal group annotations.

Sequence and algorithm choice

When attempting to find orthologous genes, the most important choice is the type of sequence to use in the comparison: amino acid or nucleic acid. Protein sequences have the advantage of being fewer in number, but they can require a well annotated genomic sequence, which may not be available for all organisms. By focusing on the transcriptome, as represented by UniGene clusters, an incompletely sequenced organism

such as *Xenopus laevis* can be included. Additionally, gene expression microarray probes are frequently designed using UniGene clusters as a template. By using UniGene as the basis for orthology comparison, direct use in microarray analysis is easier.

The choice of the comparison algorithm is the next most important decision in determining the quality of calculated orthologies. As the evolutionary distance between organisms increases, the variation between nucleic acid sequences also increases; however, because some variations in nucleic acid sequence do not affect the amino acid encoded, protein sequences generally have fewer variations than nucleic acid sequences. Also, the directionality of nucleic acid sequences in the database is not always known or correct. For both of these reasons, the TBLASTX algorithm was used. TBLASTX takes as input DNA sequences and converts them to amino acid sequences in all open reading frames (ORFs) in both the 5' and 3' directions, yielding 6 possible ORFs.

Reciprocal group composition

One issue with using UniGene clusters as the main source for transcript sequences is that due to the clustering algorithm, one can't be sure that a single UniGene cluster represents a given gene. This is because UniGene clusters can include multiple splice variants for the same gene; as an unsupervised algorithm, there is no guarantee that multiple variants will converge into a single cluster. Additionally, for some organisms, such as the pseudotetraploid *Xenopus laevis*, there can be multiple (possibly degenerate) copies of a gene present in the genome. The issue of a gene being represented by multiple transcripts in the database is addressed in two ways. First, each transcript is compared to the sequences from the same organism using the BLASTN algorithm. While this lets one see

if a sequence is unique within the same organism, these results are not used in the creation of reciprocal groups. When creating a reciprocal group network, only inter-species matches are followed, but there is no limit placed on the number of transcripts that an organism could contribute to a network. For example, the *Xenopus laevis* gene *Nkx2-5* has two independent alleles in the frog genome: Xl.22859 and Xl.48440. The best-match reciprocal group for this gene contains both alleles because the human, zebrafish, and mouse *Nkx2-5* genes form a reciprocal best-match with Xl.22859 and the rat and chicken *Nkx2-5* genes form a reciprocal best-match with Xl.48440. By not limiting the reciprocal groups to one member per organism, both copies of this allele are successfully captured in the same group (Figure 4.6). CrossGene does not explicitly handle paralogs. Because of this, it relies on having multiple orthologs in order to add paralogs to a network. This may not always be successful and is one area that needs further development.

This flexibility does have a potential downside; the reciprocal best-match group may sometimes capture a closely related gene family. For example, the reciprocal best-match group for the *CHN1* gene also contains the *CHN2* gene (Figure 4.8). The graph shows two separate but connected sub-graphs for the *CHN1* and *CHN2* genes. These genes are highly related and share connections through two *Caenorhabditis elegans* and *D. melanoster* genes: *chin-1* (Cel.29497) and *RhoGAP5A* (Dm.4631). The human *CHN1* gene is more closely related to mouse *Chn2* than to mouse *Chn1*, which further connects the two gene groups. For this gene group, the best-match and high-quality reciprocal groups (Figure 4.9) are largely identical, with the addition of the mouse *Chn1* gene (Mm.476833) which only had high-quality reciprocal matches. Interestingly, the

transcripts annotated as human *CHN2* (Hs.663145, Hs.654753, and Hs.654611) are also missing from both the best-match and high-quality reciprocal groups, as they formed no reciprocal matches to any other *CHN1*/*CHN2* gene. The human *CHN1* and *CHN2* genes are not very similar to each other when compared using the BLASTN algorithm. The human *CHN2* is similar to the mouse *Chn2* gene (e-value 1e-113), but this value was below the threshold to be called a high-quality match. Even though the composition of the best-match and high-quality reciprocal groups for *CHN1*/*CHN2* are almost identical, the high-quality group is much more interconnected.

Another downside with the flexibility of unbounded reciprocal group assembly is that the number of members in a network is highly variable. The largest reciprocal best match group contains 97 members (Figure 4.10), and contains a variety of coagulation factors and proteases linked together by a small number of *Caenorhabditis elegans* and *Drosophila melanogaster* proteases. When the *Caenorhabditis elegans* genes are removed, the group is only split into 3 subgroups with sizes of 5, 24, and 64 with 2 orphans (data not shown). However, when both the *Caenorhabditis elegans* (n=2) and *Drosophila melanogaster* (n=11) genes are removed, 13 separate subgroups are formed (Figure 4.11), ranging in size from 3 to 10 members, with 2 orphaned genes. In this extreme case, the underlying concept of blindly merging GO term annotations from the whole reciprocal group can break down, as it is highly unlikely that all the members of the group should share all annotations. However, out of 22,457 best-match groups with more than one member, only 270 (1.2%) have more than 20 members, so this problem may be restricted to a small subset of reciprocal groups. Inspection of these graphs would

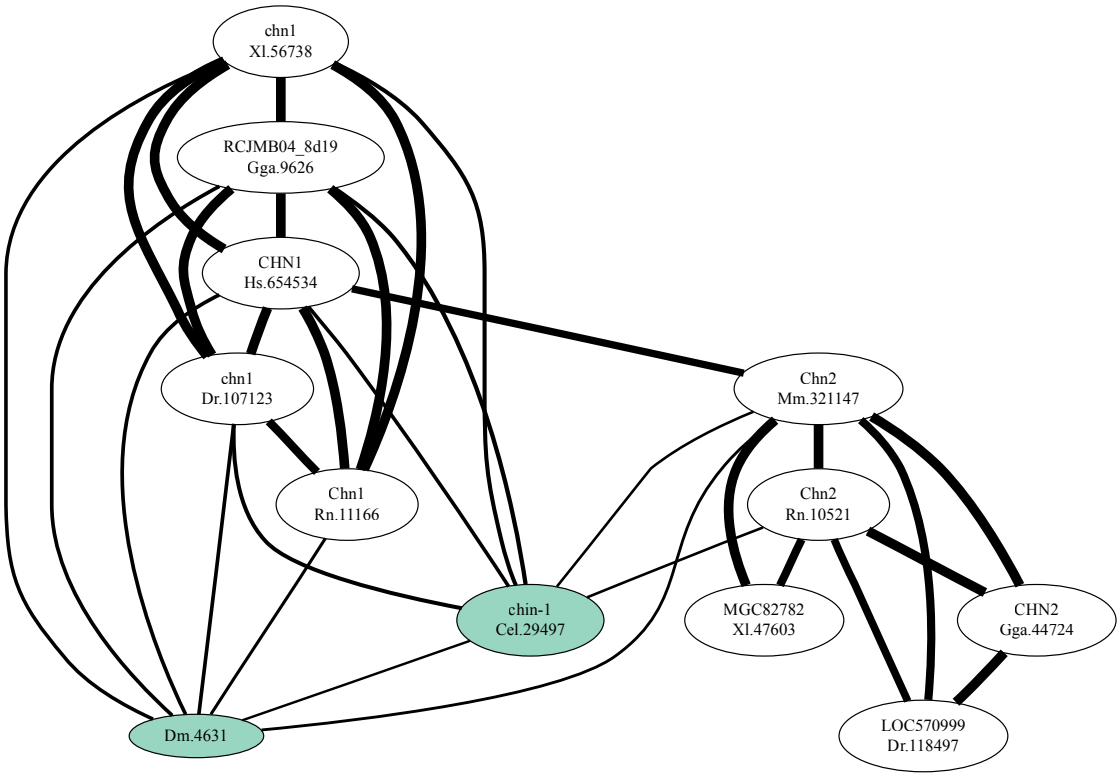


Figure 4.8 – Reciprocal best-match group for CHN1/CHN2

The network is separated into two distinct groups: CHN1 and CHN2. The two groups are bridged by the *Caenorhabditis elegans* and *Drosophila melanogaster* ancestor genes *chin-1* (Cel.29497) and *RhoGAP5A* (Dm.4631), respectively (green). The groups are also connected via a reciprocal best-match between human *CHN1* and mouse *Chn2*. The widths of the edges are inversely proportional to the e-value of the match.

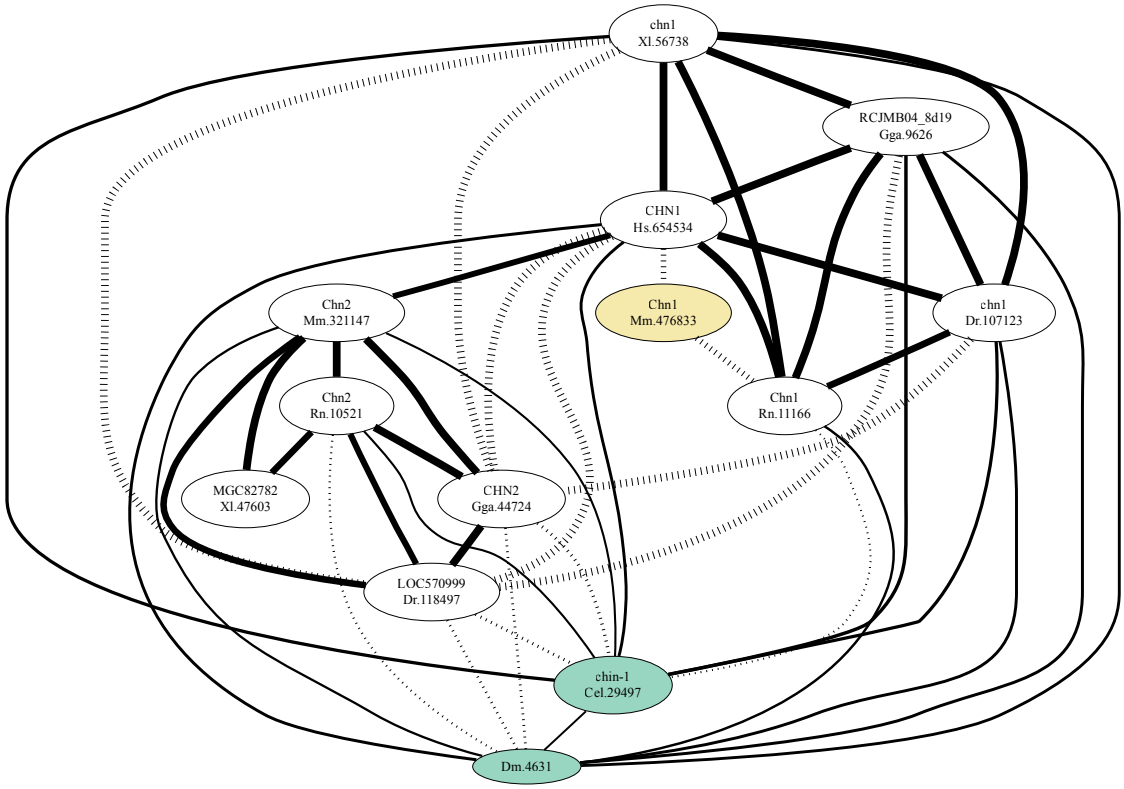


Figure 4.9 – High-quality reciprocal group for CHN1/CHN2

The composition of the two groups is identical to Figure 4.8, with the addition of the mouse *Chn1* gene (Mm.476833) (yellow), which is only a reciprocal high-quality match for human and rat *CHN1*. By including edges for high-quality matches the GraphViz dot algorithm produces a very different graph layout, which is still largely divided into two distinct *CHN1* and *CHN2* groups that are now more interconnected. The widths of the edges are inversely proportional to the e-value of the match. Best matches are shown with solid lines, high-quality matches are shown in dashed lines.

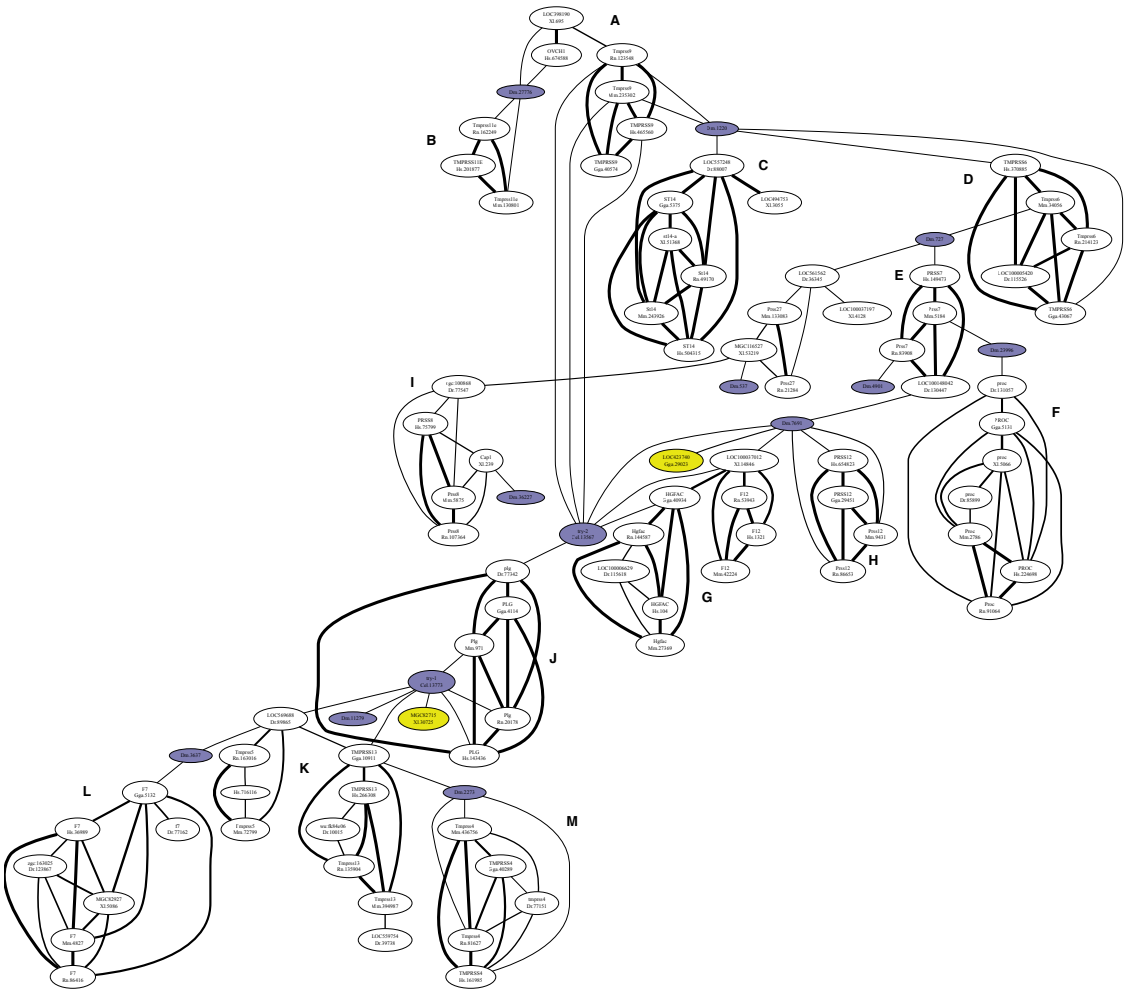


Figure 4.10 – Reciprocal group 654

Best-match group #654 contains 97 genes, consisting of mainly coagulation factors and proteases. The full reciprocal group is large, but it still forms a completely connected graph. *Caenorhabditis elegans* and *Drosophila melanogaster* genes (13) are highlighted in purple. When the Caenorhabditis elegans and Drosophila melanogaster genes are removed, 13 sub-graphs are formed (A-M), leaving 2 orphaned genes that are disconnected from all others (yellow).

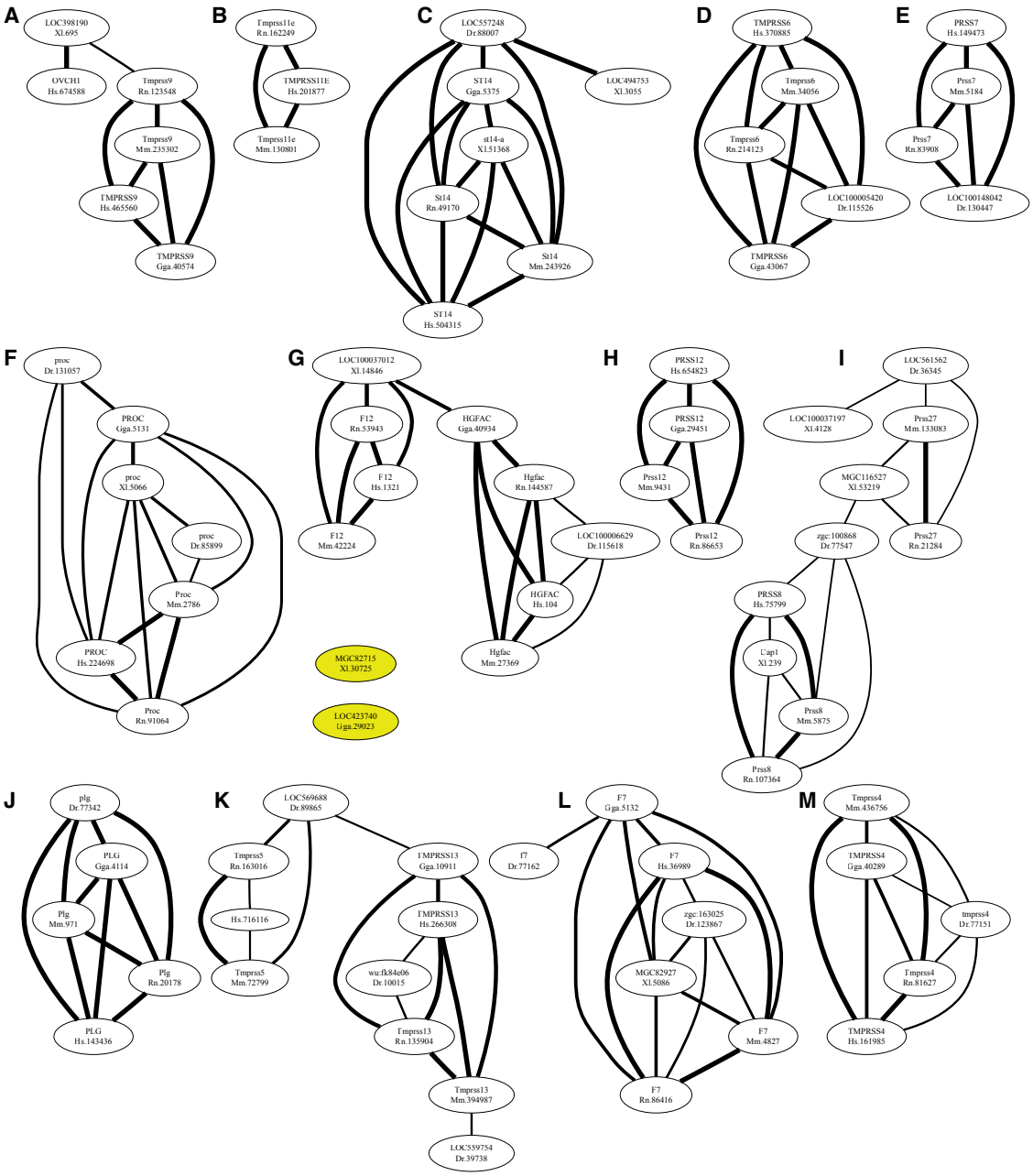


Figure 4.11 – Reciprocal group 654, trimmed

When *Caenorhabditis elegans* and *Drosophila melanogaster* genes are removed (Figure 4.10, shown in purple) from the graph, 13 independent groups are left ranging in size from 3 to 10 (A-M), leaving 2 orphaned genes that are disconnected from all others (yellow).

allow an investigator to focus on the most related subgroups and decide if manually omitting a connection would clarify the annotations.

Reciprocal group GO annotation

Best-match reciprocal groups provide a good estimate of orthologous genes among multiple organisms. This relationship is exploited to enhance the annotation of all members of an orthologous group. For example, *Nkx2-5* is a cardiogenic transcription factor involved in early determination of the heart field in the developing embryo, and its role in embryonic and adult hearts has been well established in a variety of organisms (Komuro et al. 1993; Lints et al. 1993; Grow et al. 1998). However, in the UniProt GOA database, the *Xenopus laevis Nkx2-5* gene (UniProt: P42583 and Q7T0T3) lacks the GO term annotation for heart development (GO:0007507). In the same database, the *Mus musculus Nkx2-5* gene (UniProt:P42582,Q3UQU2,Q925V3) is correctly annotated with the GO term for heart development. In fact, *Mus musculus* is the only organism whose *Nkx2-5* gene is annotated with the heart development GO term using a non-IEA evidence code (Figure 4.5). By using the collective annotation of the reciprocal best-match group, this critical annotation is now correctly applied to the *Xenopus laevis Nkx2-5* gene.

Conclusions

The functional annotation of any one organism contains many gaps. However, when combined with the annotations from other organisms, those gaps can be reduced. We used sequence similarity between transcripts to predict orthologs. By using reciprocal best-match relationships, CrossGene was able to greatly augment the annotation of a gene, based upon the existing annotations of orthologs. The increased annotation was

most striking in less annotated organisms, but was substantial even in well-studied ones. CrossGene greatly augments the number of GO annotations even for human transcripts, which showed a 46% increase in average number of annotations. Additionally, the lack of redundancy in GO annotations argues for strategies like CrossGene in order to take full advantage of cross-species annotation. In providing a more complete annotation, CrossGene can help researchers better understand the functions of unknown genes in the context of genomic scale experiments.

CHAPTER 5: CONCLUSIONS

Genomics and microarray technology have changed the way in which molecular biology is studied. We no longer only study the effect that perturbing one gene may have on another; rather, we can now study broad changes in expression across the entire transcriptome. This study aimed to exploit microarray technology to find novel targets of Nkx2-5 in early cardiogenesis, and in doing so, learn how to improve the use of *Xenopus laevis* as a model system by better annotating its genes.

It was shown that over-expressing Nkx2-5 in *Xenopus laevis* embryos caused broad changes in the expression of genes involved in development. The GO annotations of genes that were differentially expressed were enriched for transcriptional activity and DNA binding (Table 2.2). Moreover, developmental genes tended to be up-regulated, confirming the role of Nkx2-5 as positive regulator of development (Table 2.3). Using the CrossGene database (Chapter 4) to map *Xenopus laevis* genes to *Mus musculus* genes, network analysis was performed and found that many of the affected gene networks were related to development and cardiogenesis (Figure 2.9, Figure 2.10, Table 2.6). Finally, using GO annotation and spatial gene expression patterns, a list of potential targets of Nkx2-5 was compiled (Table 2.7).

Some of these potential Nkx2-5 targets were further characterized with gene expression profiling (Chapter 3). Initially, semi-quantitative RT-PCR was used to narrow a list of candidate genes for quantitative real-time PCR profiling. Two candidate genes (Hbox1 and Zic5) that were up-regulated in the Nkx2-5 over-expression microarrays had expression patterns that correlated well with the pattern of Nkx2-5 (Figure 3.11, Table

3.6). While correlation does not necessarily mean that Nkx2-5 has a direct regulatory effect on Hbox1 or Zic5, both Hbox1 and Zic5 have potential NK2 and HBOX binding sites in their promoter regions (Table 2.7), so it is possible that Nkx2-5 does directly act on them to help regulate their expression. The correlated expression patterns of Nkx2-5, Hbox1, and Zic5 could also be indicative of a common regulatory mechanism, such as TGF-β signaling.

Down-regulated candidate genes (Mix1 and Bix2) had expression profiles that showed no correlation with Nkx2-5 (Figure 3.11, Table 3.6). This means that direct regulation of these genes by Nkx2-5 is unlikely.

The genome of *Xenopus laevis* has not been sequenced, and because of its allotetraploidy, it is not likely to be sequenced in the future. This has also hindered genetic studies to help annotate the *Xenopus laevis* genome. For GO enrichment and network analysis of the Nkx2-5 over-expression microarray data, improved annotation was needed. To help improve the annotation of *Xenopus laevis*, a new annotation database was constructed (Chapter 4). The CrossGene database was build to find orthologous genes by calculating reciprocal best BLAST matches between seven species: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. The reciprocal best BLAST matches were then assembled into networks of orthologous genes. The composition of these networks compared favorably to HomoloGene clusters, with between 60%-90% of orthologous pairs maintained in CrossGene reciprocal groups. The GO annotations for each member of the networks were then applied to all of the other members, to enhance the annotation of the entire network.

Even well-studied organisms, such as human, benefitted from cross-species annotation, resulting in 46% more annotations per gene (Table 4.5). For less well annotated organisms, in particular *Xenopus laevis*, annotation was greatly enriched, with 280% more annotations per gene.

The robustness of existing GO annotations was tested by attempting to "rescue" all previously known annotations of an organism using CrossGene reciprocal group annotations (Table 4.6, Table 4.7). Robustness of a GO annotation term is a measure of how many organisms are annotated with that term within in an ortholog group. If an annotation appears in more than one organism, it is robust and could be "rescued" using the annotations of the other members of the ortholog group. If a term only appears in one organism, it cannot be "rescued" and is not robust. For all organisms, existing GO annotations lacked robustness, especially when the rescue required the more stringent non-IEA (non-computational) GO annotations. This illustrates that even for well-studied organisms, there are still gaps and missing associations in the primary GO annotation databases. This emphasizes the need for computational forms of annotation, such as CrossGene, to provide a more complete picture of a gene's function.

A concrete example of this problem is Nkx2-5. Nkx2-5 is annotated as having a role in heart development (GO:0007507) only in mouse. For human, rat, and chicken, this annotation is only covered using non-experimental or IEA (computationally derived annotation) evidence. The role of Nkx2-5 in heart development is a well-known activity that is supported by the literature in multiple organisms, but is not annotated as such in the EBI Gene Ontology Annotation database (Camon et al. 2004b). If a researcher were to use only non-IEA annotations in their analysis, this well-known function of Nkx2-5

would be missed. This argues strongly for using methods, like CrossGene, that combine known annotations from multiple organisms for a more thorough functional annotation. It also emphasizes the need for including IEA GO annotation evidence in downstream analysis.

After the Nkx2-5 microarray experiment was completed, candidate genes for RT-PCR and qPCR expression profiling were selected after the first version of the CrossGene database was completed. It wasn't until after expression profiling was completed that the CrossGene database took its current form and the more thorough analysis of changes in expression (Chapter 2) was possible. The order in which the work was completed was a necessary evil, and it made the analysis less cohesive. One example of this is that the potential targets identified in Chapter 2 after CrossGene was completed were not necessarily the same as those that were profiled (Chapter 3). While there was a good rationale behind choosing the genes that were profiled, it is clear from the analysis in Chapter 2 that certain obvious choices were excluded in these profiles. Additionally, there was one gene that was initially selected for qPCR profiling that didn't make the list of 99 putative targets in Chapter 2 (Tie2). Also, several of the potential targets in Chapter 2 would have made more compelling candidates for profiling, such as Twist and Pax3. However, this is a risk whenever data are re-analyzed using updated information.

One of the interesting results was the combination of Nkx2-5 and NF-κB in an Ingenuity derived network (Figure 2.9). NF-κB is a protein complex that has a significant role in the inflammatory response (Sen et al. 1986; Barnes et al. 1997) and apoptosis (Barkett et al. 1999). Controlled cell death, through apoptosis, is an important part of an organism's

development (Meier et al. 2000), but this only occurs much later in development when it is necessary for the organism to remove extraneous tissue. NF-κB is also known to have a role in mediating the adult response to myocardial diseases, but this doesn't necessarily extend to the developing heart (Haudek et al. 2001; Carlson et al. 2003). While NF-κB has no known role in cardiovascular development (Oka et al. 2007), it is possible that Nkx2-5 and NF-κB could indirectly interact through the TGF-β signaling pathway (Bitzer et al. 2000). Since the expression of Nkx2-5 is partially driven through the TGF-β pathway, it is possible that there is some amount of crosstalk between the two pathways through common SMAD factors. Another possible link between the two is through Meox1 and Meox2 (Figure 2.9). Like Nkx2-5, Meox1 and Meox2 are both homeodomain transcription factors that are involved in myogenesis (Petropoulos et al. 2004). Additionally, Meox2 is involved in the development of cardiomyocytes in mouse embryos (Skopicki et al. 1997) and has been shown to interact with NF-κB to inhibit angiogenesis (Patel et al. 2005). Additionally, Meox1 and Meox2 have also been known to associate with Pax3 (Stamataki et al. 2001), which was one of the predicted Nkx2-5 targets that matched all three of the prioritization criteria in Chapter 2. Evidence for a link between Nkx2-5 and NF-κB is weak, but it does underscore the potential for crosstalk between different signaling pathways in early development.

One explanation for the broad set of changes in genes related to development (Chapter 2) is that by injecting Nkx2-5 RNA, the timing of development was accelerated. Development is a complex series of events that are well coordinated. Any changes in that timing could have profound effects on many different pathways and genes. This could be accomplished through changes in the expression of common developmental regulators,

such as BMP or sonic hedgehog (Shh) signaling. Both of these genes were up-regulated in response to Nkx2-5 overexpression. Like the potential NF-κB interactions, this would not need to be through a direct interaction, but could be the result of molecular cross-talk between pathways. When measured against a control at a single time point, any acceleration in development timing would be seen as a broad up-regulation of developmentally related genes. This is one possible interpretation of the GO enrichment analysis in Chapter 2 and is one potential mechanism by which a tissue specific transcription factor, such as Nkx2-5, could have had such a large effect on so many developmental genes.

One method to test this possibility would be to re-examine the changes in gene expression caused by Nkx2-5 overexpression using a series of time points. Again using microarrays, multiple development stages could be used for gene expression profiling. This way, any changes in the expression patterns of genes could be compared not only in terms of scale (up or down regulated) but also in terms of timing. In this case, changes in the expression pattern of a gene would be far more informative than expression level differences at any single time point.

Future studies should also focus on confirmation of the targets predicted in Chapter 2. This could be done using a variety of techniques. A low-throughput technique would be ChIP assay. This technique involves using an immunoprecipitation to pull down a protein bound to DNA/chromatin. Next, using targeted primers, PCR is used to confirm that a protein was in fact bound to a specific promoter. The prospect of this type of assay in the future was the reason for using Nkx2-5 tagged with an HA epitope. Unfortunately, a

ChIP assay requires knowing the DNA sequence of the promoter for the targeted gene, so that PCR primers can be properly designed to amplify the bound region. It is not known if the *Xenopus tropicalis* genome sequence is similar enough to *Xenopus laevis* genome to conduct this experiment in *Xenopus laevis*. A high-throughput approach would be a ChIP-seq assay, which starts with the same immunoprecipitation, but instead of using promoter specific PCR primers, all genomic positions where Nkx2-5 was bound are be determined using next-generation sequencing. In this case, it would be easier to use the *Xenopus tropicalis* genome as a surrogate for *Xenopus laevis*, but there may still be other issues mapping next-generation sequencing reads from *Xenopus laevis* to the *Xenopus tropicalis* genome. While it is possible to for *Xenopus tropicalis* mRNA to hybridize to a *Xenopus laevis* cDNA microarray (Figure 1.3), it isn't known how homologous their promoters are. Because promoter regions aren't under as much pressure to be conserved evolutionarily, they can be more divergent than protein coding genes, which would make mapping more difficult. Additionally, because of the allotetraploid nature of the *Xenopus laevis* genome, multiple loci from the *Xenopus laevis* genome could map to a single position on the *Xenopus tropicalis* genome. This would make data processing much more difficult.

Because of the difficulty in dealing with the limitations of the unsequenced *Xenopus laevis* genome, I believe that any further follow-up confirmation may be better suited to be performed in a different model system. The murine P19CL6 cardiomyocyte induction cell line (Chapter 1) or *Xenopus tropicalis* would be good systems for this. If these results in the P19CL6 cell induction model could be reliably replicated, this would be my preferred system for follow-up work, due to the extensive annotation of the mouse

genome and maturity of the assembled mouse genomic sequence. Additionally, at this point, the P19CL6 cardiomyocyte induction model is the most common system for studying cardiomyogenesis. At this point in time, the state of *Xenopus tropicalis* genomic sequence is still quite immature, and so mouse would be my preferred model. Switching to the P19CL6 induction model would be a big change, so in this case, the microarray experiment presented here should be replicated in that system.

In addition to updating the database with new UniGene builds, future work on the CrossGene database should attempt to resolve the issue of properly annotating excessively large reciprocal networks. One method would be to simply break apart excessively large networks into smaller sub-networks by removing nodes that have a high "betweenness centrality" (Freeman 1977). This is a measure of how important a particular node is in linking the various parts of a graph. In CrossGene reciprocal group networks, ancestor genes that link multiple sub-networks have a high betweenness centrality. If these were simply removed from the larger networks, or made to be members of all of the sub-networks without linking the sub-networks, this could reduce the problem of overly large networks and over annotation they cause.

Greater insight into the molecular signaling in early cardiogenesis would increase our understanding of the mechanisms behind many congenital heart diseases. This in turn could enable further treatment options. Additionally, characterizing the signaling pathways that drive differentiation of pluripotent cells into cardiomyocytes could aide in the development of therapeutic treatment options for adults suffering from heart disease or who have had myocardial infarction. The potential targets identified in this study represent a solid first step in that direction.

APPENDIX 1: PCR PRIMERS

Table A1.1 – Primer3 design parameters

Primers were designed with Primer3 software using these parameters (Rozen et al. 2000).

Parameter	Value
Optimal size	24 bp
Minimum size	22 bp
Maximum size	25 bp
GC Clamp	No
Optimal Tm	60
Minimum Tm	58
Maximum Tm	64

Table A1.2 – Primer sequences used in this study

Target/reference	Sequence (Forward/Reverse)	Product size (bp)
Injection confirmation		
ODC-RT	ATCGTATCGTAGAAAGGTTTGAGC AGATCTGGTACTTCAGGGAGAATG	294
Nkx2-5HA	CTTACAATTCCCCATACAATGTCA TGGTAACCAGATCCTAGTCAGTCA	283
Nkx2-5-RT	AGATGTCTACTGAAGCACTGATGC GTTATCATTTTGATCAGGGAAACC	281
Semi-quantitative RT-PCR profiling		
AFFX-Xl-a1Act-3_s_at	GAGAGGTATCCTGACCCTGAAGTA TATATGTTGCTTGGAGGAGTGTGT	977
AFFX-Xl-bAct-5_s_at	GCCAATATATGAAGGCTATGCTCT TCCAGACAGAGTATTTACGCTCAG	533
Xl.580.1.S1_at	AAAACTGAACCGGTAAAACTTGAG CCATTTATTTCTGGTGGTGTCATA	626
Xl.6393.1.S1_at	AAGATCCTTCAGGCATTATTTCAG CCCACATCTGTCACATATTTCATT	632
Xl.793.1.A1_at	CATGAGGAGAAGGAATGTGTAGTG ACTTGCTAGACATTTTTCGGTTTC	580
Xl.824.1.S1_at	AAACCAGAGGTGTATTCTACCAGC TTTTGTGTCTGAACCATTGTCTTT	677
Xl.481.1.S1_at	GGGAAGATACCATTATGACACACA TGTTGATATAGGCAACCAGTGAGT	575
Xl.975.1.S1_at	ATATTTTTCCTAGGCCTTCTGCTT GTGCAGAACTGAGATATTTGGATG	756
Xl.397.1.S1_at	CTTCTCCTCCAGTGACCATCTAAT CACATAGTTCTCCCTCAATCTCCT	878
Xl.18750.1.S1_at	GTATCCAAACCCAGAATCCACTAC ACAATGTTCTCCATCTTCCTCTTC	900
Xl.1685.1.S1_at	GACAGTTCTCAAATCCCTGTTCTT ACTTTAGCATACACCTCCGCTATC	544

Table A1.2 (continued)

Target/reference	Sequence (Forward/Reverse)	Product size (bp)
Xl.21901.1.S1_at	CCGTATTCTTGTGGACTTACACTG AGATTCAAACGTCATTCTCAAACA	639
Xl.866.1.S2_at	TCATTTATACAGCAGATCCCAAGA ATATAATTTATTGAACCTCGCCCA	1008
Xl.12160.1.S1_at	GAAAAGACTGAAAGTGCAACTCAA AAATATCTCAACTTCTGGTCTGCC	541
Xl.1093.1.S1_at	CTTACAATTCCCCATACAATGTCA AACTGGCAGTATAAGGCACATACA	1057
Xl.16169.1.S1_at	CATGTGAATGTGGCATTTTTATTT TATACACAAGCTTAGCGCTGTTTC	800
Xl.823.1.S1_at	AGAGAAGGAGGCTCAGTAAAGTCA TATATAAAGGAGCATCTGGCACAA	943
Xl.1043.1.S1_at	ACAAAAGAAGTTAAAGAGCGCCTA TTTAGAGCAGGGCTTTTTAGAAGA	884
Xl.933.1.S1_at	AAGTATGTCAATGGAGAATGGGTT AAGAGAAGTTGAGCTCCGAAGATA	846
Xl.8190.1.S1_at	AAGCCATCATTATTCTAGCACCTC AAACATTCTCTTCCCAGTCTGAAC	605
Xl.146.1.S1_at	AAGAGATTTCCTGAAACCTGATTG TATAACACATGGTAGATTGGTCGG	833
Xl.13437.1.A1_at	GGAAGGTCATTCAATCTCTTCTGT ATATTTCTTGCTGCTAAATCCCTG	924
Xl.1394.1.S1_at	GATGATGAGACAGCCATTAATCAG TTATCATGTATTTTTAGAGCGCCA	863
Xl.1209.1.S1_at	GAGAGAGACAGAGAGAGAGCCAGT AGGAATAACGCAGAGACTGAGAGT	903
Xl.8333.1.A1_at	AGTCCATGTCTTTGTGACAGTGTT GTATTTGACTCCAGCAGAGCACTA	637
Xl.25598.1.S1_at	ATGGGTTTGGATCGTATAAAAGAA AAGTGTCTCACCAACATTACTCCA	565

Table A1.2 (continued)

Target/reference	Sequence (Forward/Reverse)	Product size (bp)
Xl.25289.1.A1_at	TTCATTACTTGGAGTCAGAATGGA TGAGAAATAACGCCACTTTCATTA	1004
Xl.16644.1.S1_at	ATCCCTTTTAAAGAGTACCAAGGG AGTGTTTAACAGGAAAGTTGGGAG	803
Xl.15793.1.A1_at	CAGAGATAGCTACAGCCAACTGAA CTGTATGATTGCGTAGTTTCCATC	682
Xl.15970.1.A1_at	TATGGATACTGAGGCAAAGAATCA ATGAACCCATCATTAAGGACACTT	780
Xl.9113.1.A1_at	AAGTTTACAGTTGTGACCTGACCA CATCTCACAAATGATTTCTTCCTG	618
Xl.19790.1.S1_at	GGAATCAGACAGCTACAAGCTACA GATGCTTTCATTGATGCTCATTAC	624
Xl.20765.1.A1_at	ATATTAATACCCAGCAGAGCTTCG CATCCTCTATTAGGGACTGCTGAT	645
Xl.13885.1.A1_s_at	AATGGGACCTGACTAAACAAAGAG AATTGTTGTGGAAAGAGCACATTA	588
β-actin	GCACCAGAAGAACACCCAGT CTGGAAGAGTGCCTCTGGAC	500
Real-time qPCR		
Nkx2-5-rt	GCTCTCCTTTGAAAAGCCCT GCCTGGAAGTGATGTCCATT	196
Bix2-rt	AGGAACTGGCCAGACAAATG TGCTATGGTGATTGTGCCAT	182
CHN1-rt	CTCCAGATCCTGATGCACAA GTTGGGCCGAACACTATACC	169
HBox1-rt	TAGCTGCAGGCAGAACTCAA CCAGAGTTTGGTAACGGGAA	209
Mix1-Xanthos (Xanthos et al. 2001)	GCAGATGCCAGTTCAGCCAATG TTTGTCCATAGGTTCCGCCCTG	188

Table A1.2 (continued)

Target/reference	Sequence (Forward/Reverse)	Product size (bp)
Zic5-rt	TGGGAGTATGGGCTATCCTG CATATTAACGGTTGCCCCTG	157
ODC-Heasman (Heasman et al. 2000)	GCCATTGTGAAGACTCTCTCCATTC TTCGGGTGATTCCTTGCCAC	221
GAPDH-rt	CTTTGATGCTGATGCTGGAA GACAGACTAGCAGGATGGGC	195
Miscellaneous primers		
M13 Forward	TGTAAAACGACGGCCAGT	-
M13 Reverse	CAGGAAACAGCTATGACC	-
Anchored oligo-dT	TTTTTTTTTTTTTTTTTTV	-

APPENDIX 2: GO ENRICHMENT IN NKX2-5 OVEREXPRESSION MICROARRAYS

Table A2.1 – Biological Process – up-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0003002	regionalization	11	31	5.5E-08
GO:0007389	pattern specification process	15	37	9.7E-08
GO:0035282	segmentation	4	16	1.3E-07
GO:0030154	cell differentiation	27	53	2.6E-07
GO:0009952	anterior/posterior pattern formation	6	21	3.7E-07
GO:0032501	multicellular organismal process	55	87	4.5E-07
GO:0045893	positive regulation of transcription, DNA-dependent	13	31	8.5E-07
GO:0010557	positive regulation of macromolecule biosynthetic process	16	36	9.1E-07
GO:0051254	positive regulation of RNA metabolic process	13	31	9.4E-07
GO:0009887	organ morphogenesis	15	35	1.2E-06
GO:0001756	somitogenesis	3	12	1.6E-06
GO:0048598	embryonic morphogenesis	13	30	2.5E-06
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	10	26	2.5E-06
GO:0007275	multicellular organismal development	30	54	2.8E-06
GO:0031328	positive regulation of cellular biosynthetic process	16	36	2.9E-06
GO:0009891	positive regulation of biosynthetic process	17	36	3.1E-06
GO:0048731	system development	19	39	4.0E-06
GO:0009893	positive regulation of metabolic process	20	40	8.2E-06

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0048646	anatomical structure formation involved in morphogenesis	13	30	8.8E-06
GO:0010604	positive regulation of macromolecule metabolic process	19	39	9.4E-06
GO:0007494	midgut development	2	8	1.0E-05
GO:0048705	skeletal system morphogenesis	3	11	1.1E-05
GO:0031325	positive regulation of cellular metabolic process	20	39	1.1E-05
GO:0045935	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	15	33	1.2E-05
GO:0048704	embryonic skeletal system morphogenesis	2	9	1.3E-05
GO:0007366	periodic partitioning by pair rule gene	1	4	2.1E-05
GO:0045941	positive regulation of transcription	14	31	2.1E-05
GO:0048869	cellular developmental process	37	61	2.2E-05
GO:0010628	positive regulation of gene expression	14	31	2.2E-05
GO:0021570	rhombomere 4 development	1	3	2.2E-05
GO:0001709	cell fate determination	4	13	3.4E-05
GO:0006357	regulation of transcription from RNA polymerase II promoter	17	33	4.8E-05
GO:0001501	skeletal system development	5	15	5.7E-05
GO:0009880	embryonic pattern specification	5	15	7.5E-05
GO:0048706	embryonic skeletal system development	2	7	1.3E-04
GO:0030902	hindbrain development	2	8	1.4E-04
GO:0035239	tube morphogenesis	5	13	1.7E-04
GO:0008045	motor axon guidance	2	7	1.8E-04
GO:0048532	anatomical structure arrangement	1	3	2.2E-04

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0021612	facial nerve structural organization	1	3	2.2E-04
GO:0021604	cranial nerve structural organization	1	3	2.2E-04
GO:0021561	facial nerve development	1	3	2.2E-04
GO:0048565	gut development	3	9	2.4E-04
GO:0048741	skeletal muscle fiber development	1	5	2.4E-04
GO:0030539	male genitalia development	1	5	2.4E-04
GO:0048747	muscle fiber development	1	6	3.3E-04
GO:0021545	cranial nerve development	1	4	3.7E-04
GO:0021569	rhombomere 3 development	1	3	4.2E-04
GO:0048732	gland development	4	12	4.5E-04
GO:0048856	anatomical structure development	53	74	4.7E-04
GO:0002076	osteoblast development	1	4	5.0E-04
GO:0007365	periodic partitioning	2	7	5.3E-04
GO:0001570	vasculogenesis	2	7	5.3E-04
GO:0021953	central nervous system neuron differentiation	1	4	6.7E-04
GO:0051658	maintenance of nucleus location	1	2	8.0E-04
GO:0051594	detection of glucose	1	2	8.0E-04
GO:0045721	negative regulation of gluconeogenesis	1	2	8.0E-04
GO:0042070	maintenance of oocyte nucleus location involved in oocyte dorsal/ventral axis specification	1	2	8.0E-04
GO:0034287	detection of monosaccharide stimulus	1	2	8.0E-04
GO:0010359	regulation of anion channel activity	1	2	8.0E-04
GO:0009732	detection of hexose stimulus	1	2	8.0E-04

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0009730	detection of carbohydrate stimulus	1	2	8.0E-04
GO:0055001	muscle cell development	1	5	8.5E-04
GO:0045740	positive regulation of DNA replication	1	5	8.5E-04
GO:0060606	tube closure	1	5	1.0E-03
GO:0001843	neural tube closure	1	5	1.0E-03
GO:0051603	proteolysis involved in cellular protein catabolic process	10	1	1.0E-03
GO:0044257	cellular protein catabolic process	10	1	1.0E-03
GO:0034962	cellular biopolymer catabolic process	10	1	1.0E-03
GO:0030198	extracellular matrix organization	3	8	1.1E-03
GO:0035287	head segmentation	1	4	1.1E-03
GO:0021675	nerve development	1	4	1.1E-03
GO:0009267	cellular response to starvation	1	4	1.1E-03
GO:0021754	facial nucleus development	1	3	1.1E-03
GO:0022610	biological adhesion	11	21	1.3E-03
GO:0007155	cell adhesion	11	21	1.3E-03
GO:0007399	nervous system development	11	21	1.4E-03
GO:0055002	striated muscle cell development	1	4	1.4E-03
GO:0060537	muscle tissue development	3	8	1.4E-03
GO:0043632	modification-dependent macromolecule catabolic process	9	1	1.5E-03
GO:0019941	modification-dependent protein catabolic process	9	1	1.5E-03
GO:0045165	cell fate commitment	6	14	1.6E-03
GO:0048522	positive regulation of cellular process	36	53	1.6E-03

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0048857	neural nucleus development	1	3	1.7E-03
GO:0045821	positive regulation of glycolysis	1	3	1.7E-03
GO:0042063	gliogenesis	1	3	1.7E-03
GO:0021571	rhombomere 5 development	1	3	1.7E-03
GO:0048663	neuron fate commitment	3	8	1.8E-03
GO:0009888	tissue development	12	23	2.0E-03
GO:0034621	cellular macromolecular complex subunit organization	9	1	2.3E-03
GO:0043374	CD8-positive, alpha-beta T cell differentiation	1	2	2.4E-03
GO:0043282	pharyngeal muscle development	1	2	2.4E-03
GO:0033604	negative regulation of catecholamine secretion	1	2	2.4E-03
GO:0032811	negative regulation of epinephrine secretion	1	2	2.4E-03
GO:0014060	regulation of epinephrine secretion	1	2	2.4E-03
GO:0001823	mesonephros development	1	2	2.4E-03
GO:0044267	cellular protein metabolic process	43	26	2.5E-03
GO:0007350	blastoderm segmentation	1	4	2.6E-03
GO:0050794	regulation of cellular process	95	115	2.6E-03
GO:0050877	neurological system process	13	24	2.7E-03
GO:0035148	tube lumen formation	1	5	2.8E-03
GO:0000904	cell morphogenesis involved in differentiation	3	8	2.9E-03
GO:0048513	organ development	39	55	2.9E-03
GO:0003008	system process	18	30	3.0E-03
GO:0048754	branching morphogenesis of a tube	4	9	3.0E-03

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0014029	neural crest formation	1	3	3.1E-03
GO:0042552	myelination	2	5	3.2E-03
GO:0007379	segment specification	2	5	3.2E-03
GO:0006275	regulation of DNA replication	2	7	3.2E-03
GO:0050890	cognition	12	22	3.3E-03
GO:0001658	branching involved in ureteric bud morphogenesis	1	4	3.6E-03
GO:0042475	odontogenesis of dentine-containing tooth	2	6	3.6E-03
GO:0007600	sensory perception	9	18	3.8E-03
GO:0014706	striated muscle tissue development	3	7	3.8E-03
GO:0035290	trunk segmentation	1	3	4.1E-03
GO:0030199	collagen fibril organization	1	3	4.1E-03
GO:0010556	regulation of macromolecule biosynthetic process	44	60	4.5E-03
GO:0060395	SMAD protein signal transduction	1	2	4.6E-03
GO:0051657	maintenance of organelle location	1	2	4.6E-03
GO:0051365	cellular response to potassium ion starvation	1	2	4.6E-03
GO:0051156	glucose 6-phosphate metabolic process	1	2	4.6E-03
GO:0045084	positive regulation of interleukin-12 biosynthetic process	1	2	4.6E-03
GO:0045075	regulation of interleukin-12 biosynthetic process	1	2	4.6E-03
GO:0048066	pigmentation during development	2	6	4.8E-03
GO:0008284	positive regulation of cell proliferation	8	16	4.8E-03
GO:0034622	cellular macromolecular complex assembly	8	1	4.8E-03

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0043933	macromolecular complex subunit organization	16	6	4.9E-03
GO:0046632	alpha-beta T cell differentiation	1	3	5.2E-03
GO:0035289	posterior head segmentation	1	3	5.2E-03
GO:0045595	regulation of cell differentiation	15	26	5.3E-03
GO:0006355	regulation of transcription, DNA-dependent	35	50	5.4E-03
GO:0048468	cell development	12	22	5.5E-03
GO:0060021	palate development	1	4	5.7E-03
GO:0031669	cellular response to nutrient levels	1	4	5.7E-03
GO:0002763	positive regulation of myeloid leukocyte differentiation	1	4	5.7E-03
GO:0051276	chromosome organization	10	2	6.2E-03
GO:0006576	biogenic amine metabolic process	2	6	6.2E-03
GO:0051252	regulation of RNA metabolic process	37	52	6.3E-03
GO:0001822	kidney development	3	8	6.4E-03
GO:0030097	hemopoiesis	4	9	6.5E-03
GO:0030947	regulation of vascular endothelial growth factor receptor signaling pathway	1	3	6.5E-03
GO:0021546	rhombomere development	1	3	6.5E-03
GO:0006110	regulation of glycolysis	1	3	6.5E-03
GO:0019538	protein metabolic process	49	33	6.5E-03
GO:0016337	cell-cell adhesion	5	10	6.6E-03
GO:0031326	regulation of cellular biosynthetic process	45	61	6.6E-03
GO:0042594	response to starvation	1	4	6.6E-03
GO:0018958	phenol metabolic process	1	4	6.6E-03

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0006584	catecholamine metabolic process	1	4	6.6E-03
GO:0001656	metanephros development	1	4	6.6E-03
GO:0009889	regulation of biosynthetic process	45	61	6.8E-03
GO:0031668	cellular response to extracellular stimulus	2	5	7.0E-03
GO:0008366	axon ensheathment	2	5	7.0E-03
GO:0007272	ensheathment of neurons	2	5	7.0E-03
GO:0042127	regulation of cell proliferation	16	27	7.0E-03
GO:0042476	odontogenesis	2	6	7.3E-03
GO:0045657	positive regulation of monocyte differentiation	1	2	7.6E-03
GO:0045655	regulation of monocyte differentiation	1	2	7.6E-03
GO:0042415	norepinephrine metabolic process	1	2	7.6E-03
GO:0032288	myelin assembly	1	2	7.6E-03
GO:0008359	regulation of bicoid mRNA localization	1	2	7.6E-03
GO:0006003	fructose 2,6-bisphosphate metabolic process	1	2	7.6E-03
GO:0001763	morphogenesis of a branching structure	4	9	7.6E-03
GO:0006996	organelle organization	33	19	7.7E-03
GO:0007565	female pregnancy	2	5	7.7E-03
GO:0003007	heart morphogenesis	3	7	7.8E-03
GO:0046529	imaginal disc fusion, thorax closure	1	3	7.9E-03
GO:0045596	negative regulation of cell differentiation	7	14	7.9E-03
GO:0006928	cell motion	17	27	7.9E-03
GO:0030217	T cell differentiation	2	5	8.5E-03
GO:0016202	regulation of striated muscle development	2	6	8.5E-03

Table A2.1 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0008283	cell proliferation	12	21	9.2E-03
GO:0048771	tissue remodeling	2	5	9.3E-03
GO:0048660	regulation of smooth muscle cell proliferation	2	5	9.3E-03
GO:0046631	alpha-beta T cell activation	1	3	9.6E-03
GO:0018210	peptidyl-threonine modification	1	3	9.6E-03
GO:0008103	oocyte microtubule cytoskeleton polarization	1	3	9.6E-03
GO:0010810	regulation of cell-substrate adhesion	1	4	9.7E-03
GO:0006396	RNA processing	13	4	9.7E-03
GO:0050953	sensory perception of light stimulus	5	10	1.0E-02
GO:0007601	visual perception	5	10	1.0E-02
GO:0009612	response to mechanical stimulus	3	7	1.0E-02
GO:0045639	positive regulation of myeloid cell differentiation	2	5	1.0E-02

Table A2.2 – Biological Process – down-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0048518	positive regulation of biological process	56	31	2.6E-05
GO:0040008	regulation of growth	33	14	9.5E-05
GO:0009653	anatomical structure morphogenesis	38	18	1.0E-04
GO:0016254	preassembly of GPI anchor in ER membrane	1	4	1.2E-04
GO:0045927	positive regulation of growth	28	11	1.3E-04
GO:0040009	regulation of growth rate	25	9	2.1E-04
GO:0040010	positive regulation of growth rate	25	9	2.2E-04
GO:0002119	nematode larval development	26	10	2.8E-04
GO:0008152	metabolic process	104	82	5.6E-04
GO:0032502	developmental process	87	65	6.6E-04
GO:0002164	larval development	26	11	6.8E-04
GO:0050896	response to stimulus	49	30	9.6E-04
GO:0044238	primary metabolic process	96	75	1.4E-03
GO:0042592	homeostatic process	13	3	1.6E-03
GO:0006357	regulation of transcription from RNA polymerase II promoter	15	4	1.8E-03
GO:0043170	macromolecule metabolic process	82	62	2.0E-03
GO:0044249	cellular biosynthetic process	45	28	2.5E-03
GO:0032268	regulation of cellular protein metabolic process	11	2	2.9E-03
GO:0046489	phosphoinositide biosynthetic process	1	4	2.9E-03
GO:0000003	reproduction	24	11	3.2E-03
GO:0051246	regulation of protein metabolic process	12	3	3.3E-03
GO:0045595	regulation of cell differentiation	14	4	3.5E-03

Table A2.2 (continued)

GO	Name	Expected	Actual	Fisher p-value
GO:0065008	regulation of biological quality	25	12	3.8E-03
GO:0006506	GPI anchor biosynthetic process	1	3	3.9E-03
GO:0006505	GPI anchor metabolic process	1	3	3.9E-03
GO:0009058	biosynthetic process	48	31	5.0E-03
GO:0043283	biopolymer metabolic process	80	62	5.7E-03
GO:0042221	response to chemical stimulus	26	13	5.9E-03
GO:0051239	regulation of multicellular organismal process	28	15	5.9E-03
GO:0065007	biological regulation	109	91	6.0E-03
GO:0048519	negative regulation of biological process	38	23	6.1E-03
GO:0030325	adrenal gland development	1	2	6.2E-03
GO:0007610	behavior	14	5	6.8E-03
GO:0050789	regulation of biological process	104	87	7.1E-03
GO:0007049	cell cycle	13	23	7.2E-03
GO:0048513	organ development	35	21	8.5E-03
GO:0040007	growth	26	14	8.9E-03
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	49	33	9.0E-03
GO:0080090	regulation of primary metabolic process	48	33	9.0E-03
GO:0017015	regulation of transforming growth factor beta receptor signaling pathway	2	6	9.3E-03
GO:0009059	macromolecule biosynthetic process	37	23	1.0E-02
GO:0009791	post-embryonic development	27	15	1.0E-02

Table A2.3 – Molecular Function – up-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0003700	transcription factor activity	17	33	1.6E-04
GO:0043565	sequence-specific DNA binding	13	26	5.8E-04
GO:0003705	RNA polymerase II transcription factor activity, enhancer binding	2	7	7.1E-04
GO:0004396	hexokinase activity	1	2	8.1E-04
GO:0004340	glucokinase activity	1	2	8.1E-04
GO:0005100	Rho GTPase activator activity	1	4	9.0E-04
GO:0016563	transcription activator activity	11	21	2.1E-03
GO:0030528	transcription regulator activity	28	43	2.3E-03
GO:0003677	DNA binding	33	49	3.4E-03
GO:0003702	RNA polymerase II transcription factor activity	7	15	3.5E-03
GO:0046982	protein heterodimerization activity	6	13	4.3E-03
GO:0005099	Ras GTPase activator activity	2	5	6.5E-03
GO:0043138	3'-5' DNA helicase activity	1	3	8.1E-03
GO:0003723	RNA binding	16	6	9.4E-03

Table A2.4 – Molecular Function – down-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0004587	ornithine-oxo-acid transaminase activity	1	2	6.5E-04
GO:0003992	N2-acetyl-L-ornithine:2-oxoglutarate 5-aminotransferase activity	1	2	6.5E-04
GO:0030695	GTPase regulator activity	5	12	6.7E-04
GO:0060589	nucleoside-triphosphatase regulator activity	5	12	1.0E-03
GO:0017176	phosphatidylinositol N-acetylglucosaminyltransferase activity	1	2	1.9E-03
GO:0004137	deoxycytidine kinase activity	1	2	1.9E-03
GO:0005083	small GTPase regulator activity	3	9	2.8E-03
GO:0005198	structural molecule activity	11	2	2.9E-03
GO:0019957	C-C chemokine binding	1	2	3.8E-03
GO:0016494	C-X-C chemokine receptor activity	1	2	3.8E-03
GO:0016493	C-C chemokine receptor activity	1	2	3.8E-03
GO:0005089	Rho guanyl-nucleotide exchange factor activity	1	4	4.0E-03
GO:0019958	C-X-C chemokine binding	1	2	6.2E-03
GO:0019136	deoxynucleoside kinase activity	1	2	6.2E-03
GO:0004950	chemokine receptor activity	1	2	6.2E-03
GO:0001637	G-protein chemoattractant receptor activity	1	2	6.2E-03
GO:0003723	RNA binding	14	5	9.5E-03

Table A2.5 – Cellular Component – up-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0030529	ribonucleoprotein complex	13	1	6.6E-05
GO:0005576	extracellular region	17	34	7.4E-05
GO:0043218	compact myelin	1	3	4.3E-04
GO:0005622	intracellular	24	10	1.1E-03
GO:0043226	organelle	132	112	1.9E-03
GO:0044421	extracellular region part	12	23	2.0E-03
GO:0005615	extracellular space	10	20	2.5E-03
GO:0043229	intracellular organelle	132	112	2.6E-03
GO:0043232	intracellular non-membrane-bounded organelle	38	23	5.4E-03
GO:0043228	non-membrane-bounded organelle	38	23	5.4E-03
GO:0044444	cytoplasmic part	84	65	7.1E-03
GO:0031012	extracellular matrix	5	10	7.1E-03
GO:0032991	macromolecular complex	56	40	9.4E-03
GO:0005578	proteinaceous extracellular matrix	4	9	1.0E-02

Table A2.6 – Cellular Component – down-regulated genes

GO	Name	Expected	Actual	Fisher p-value
GO:0005813	centrosome	5	16	2.7E-06
GO:0005815	microtubule organizing center	5	16	1.5E-05
GO:0032991	macromolecular complex	51	31	6.4E-04
GO:0030529	ribonucleoprotein complex	12	2	1.4E-03
GO:0044430	cytoskeletal part	15	27	1.9E-03
GO:0044421	extracellular region part	11	2	2.0E-03
GO:0044428	nuclear part	37	21	2.4E-03
GO:0042995	cell projection	12	3	4.7E-03
GO:0005730	nucleolus	13	4	5.0E-03
GO:0031966	mitochondrial membrane	8	1	7.3E-03
GO:0005875	microtubule associated complex	2	6	7.5E-03
GO:0005615	extracellular space	10	2	8.6E-03
GO:0000267	cell fraction	20	9	9.6E-03
GO:0043234	protein complex	40	26	1.0E-02

REFERENCES

Akazawa, H. and I. Komuro (2005). "Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases." Pharmacology & therapeutics **107**(2): 252-268.

Altschul, S., T. Madden, et al. (1997a). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389.

Altschul, S. F., W. Gish, et al. (1990a). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Altschul, S. F., W. Gish, et al. (1990b). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, S. F., T. L. Madden, et al. (1997b). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Research **25**(17): 3389-3402.

Alwine, J. C., D. J. Kemp, et al. (1977). "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." Proceedings of the National Academy of Sciences of the United States of America **74**(12): 5350-5354.

Ariizumi, T., M. Kinoshita, et al. (2003). "Amphibian in vitro heart induction: a simple and reliable model for the study of vertebrate cardiac development." The International journal of developmental biology **47**(6): 405-410.

Ascher, D., P. F. Dubois, et al. (1999). Numerical Python. Manual.

Ashburner, M., C. A. Ball, et al. (2000a). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Ashburner, M., C. A. Ball, et al. (2000b). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics **25**(1): 25-29.

Barkett, M. and T. D. Gilmore (1999). "Control of apoptosis by Rel/NF-kappaB transcription factors." Oncogene **18**(49): 6910-6924.

Barnes, P. J. and M. Karin (1997). "Nuclear factor-kappaB: a pivotal transcription factor in chronic inflammatory diseases." The New England journal of medicine **336**(15): 1066-1071.

Basson, C. T., T. Huang, et al. (1999). "Different TBX5 interactions in heart and limb defined by Holt-Oram syndrome mutations." Proceedings of the National Academy of Sciences of the United States of America **96**(6): 2919-2924.

Benson, D. W., G. M. Silberbach, et al. (1999). "Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways." The Journal of clinical investigation **104**(11): 1567-1573.

Bitzer, M., G. von Gersdorff, et al. (2000). "A mechanism of suppression of TGF-beta/SMAD signaling by NF-kappa B/RelA." Genes & Development **14**(2): 187-197.

Bodmer, R. (1993). "The gene tinman is required for specification of the heart and visceral muscles in Drosophila." Development (Cambridge, England) **118**(3): 719-729.

Bowes, J. B., K. A. Snyder, et al. (2010). "Xenbase: gene expression and improved integration." Nucleic Acids Research **38**(Database issue): D607-612.
Brown, C. B., A. S. Boyer, et al. (1999). "Requirement of type III TGF-beta receptor for endocardial cell transformation in the heart." Science (New York, NY) **283**(5410): 2080-2082.
Brown, C. O., X. Chi, et al. (2004). "The cardiac determination factor, Nkx2-5, is activated by mutual cofactors GATA-4 and Smad1/4 via a novel upstream enhancer." The Journal of biological chemistry **279**(11): 10659-10669.
Bruneau, B. G. (2002). "Transcriptional regulation of vertebrate cardiac morphogenesis." Circulation Research **90**(5): 509-519.
Bruneau, B. G., G. Nemer, et al. (2001). "A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease." Cell **106**(6): 709-721.
Camon, E., M. Magrane, et al. (2004a). "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." Nucleic Acids Res **32**(Database issue): D262-266.
Camon, E., M. Magrane, et al. (2004b). "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." Nucleic Acids Research **32**(Database issue): D262-266.
Carlson, D. L., D. J. White, et al. (2003). "I kappa B overexpression in cardiomyocytes prevents NF-kappa B translocation and provides cardioprotection in trauma." American journal of physiology. Heart and circulatory physiology **284**(3): H804-814.
Chen, C. Y. and R. J. Schwartz (1995). "Identification of novel DNA binding targets and regulatory domains of a murine tinman homeodomain factor, nkx-2.5." The Journal of biological chemistry **270**(26): 15628-15633.
Choi, S. C., J. Yoon, et al. (2004). "5-azacytidine induces cardiac differentiation of P19 embryonic stem cells." Exp Mol Med **36**(6): 515-523.
Cleaver, O. B., K. D. Patterson, et al. (1996). "Overexpression of the tinman-related genes XNkx-2.5 and XNkx-2.3 in Xenopus embryos results in myocardial hyperplasia." Development (Cambridge, England) **122**(11): 3549-3556.
Consortium, G. O. (2004a). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Research **32**(Database issue): D258-261.
Consortium, I. H. G. S. (2004b). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-945.
Conway, S. J., D. J. Henderson, et al. (1997a). "Pax3 is required for cardiac neural crest migration in the mouse: evidence from the splotch (Sp2H) mutant." Development (Cambridge, England) **124**(2): 505-514.
Conway, S. J., D. J. Henderson, et al. (1997b). "Development of a lethal congenital heart defect in the splotch (Pax3) mutant mouse." Cardiovascular Research **36**(2): 163-173.
DeLano Scientific, L. (2009). The PyMOL Molecular Graphics System.
Deluca, T. F., I.-H. Wu, et al. (2006). "Roundup: a multi-genome repository of orthologs and evolutionary distances." Bioinformatics **22**(16): 2044-2046.
Durocher, D., F. Charron, et al. (1997). "The cardiac transcription factors Nkx2-5 and GATA-4 are mutual cofactors." The EMBO journal **16**(18): 5687-5696.

Durocher, D. and M. Nemer (1998). "Combinatorial interactions regulating cardiac transcription." Developmental Genetics **22**(3): 250-262.
Evans, S. M., W. Yan, et al. (1995). "tinman, a Drosophila homeobox gene required for heart and visceral mesoderm specification, may be represented by a family of genes in vertebrates: XNkx-2.3, a second vertebrate homologue of tinman." Development (Cambridge, England) **121**(11): 3889-3899.
Faber, J. and P. D. Nieuwkoop, Eds. (1994). Normal Table of Xenopus Laevis (Daudin). A Systematical & Chronological Survey of the Development from the Fertilized Egg till the End of Metamorphosis, Garland Science.
Fisher, S. R. A. (1970). Statistical methods for research workers. Fourteenth Edition Revised, Oliver & Boyd.
Fodor, S. P., J. L. Read, et al. (1991). "Light-directed, spatially addressable parallel chemical synthesis." Science (New York, NY) **251**(4995): 767-773.
Freeman, L. (1977). "A Set of Measures of Centrality Based on Betweenness." Sociometry **40**(1): 35-41.
Fu, Y., W. Yan, et al. (1998). "Vertebrate tinman homologues XNkx2-3 and XNkx2-5 are required for heart formation in a functionally redundant manner." Development (Cambridge, England) **125**(22): 4439-4449.
Gajewski, K., Y. Kim, et al. (1998). "Combinatorial control of Drosophila mef2 gene expression in cardiac and somatic muscle cell lineages." Development genes and evolution **208**(7): 382-392.
Galvin, K. M., M. J. Donovan, et al. (2000). "A role for smad6 in development and homeostasis of the cardiovascular system." Nature genetics **24**(2): 171-174.
Gansner, E. R. and Y. Koren (2007). "Improved circular layouts." Lecture Notes in Computer Science.
Gansner, E. R. and S. C. North (2000). "An open graph visualization system and its applications to software engineering." Software Practice and Experience.
Grow, M. W. and P. A. Krieg (1998). "Tinman function is essential for vertebrate heart development: elimination of cardiac differentiation by dominant inhibitory mutants of the tinman-related genes, XNkx2-3 and XNkx2-5." Developmental Biology **204**(1): 187-196.
Gruschus, J. M., D. H. Tsao, et al. (1997). "Interactions of the vnd/NK-2 homeodomain with DNA by nuclear magnetic resonance spectroscopy: basis of binding specificity." Biochemistry **36**(18): 5372-5380.
Habara-Ohkubo, A. (1996). "Differentiation of beating cardiac muscle cells from a derivative of P19 embryonal carcinoma cells." Cell structure and function **21**(2): 101-110.
Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Research **32**(Database issue): D258-261.
Harvey, R. P. (1996). "NK-2 homeobox genes and heart development." Developmental Biology **178**(2): 203-216.
Harvey, R. P., D. Lai, et al. (2002). "Homeodomain factor Nkx2-5 in heart development and disease." Cold Spring Harbor symposia on quantitative biology **67**: 107-114.
Haudek, S. B., E. Spencer, et al. (2001). "Overexpression of cardiac I-kappaBalpha prevents endotoxin-induced myocardial dysfunction." American journal of physiology. Heart and circulatory physiology **280**(3): H962-968.

Heasman, J., M. Kofron, et al. (2000). "Beta-catenin signaling activity dissected in the early Xenopus embryo: a novel antisense approach." Developmental Biology **222**(1): 124-134.
Heid, C. A., J. Stevens, et al. (1996). "Real time quantitative PCR." Genome research **6**(10): 986-994.
Hellsten, U., R. M Harland, et al. (2010). "The genome of the Western clawed frog Xenopus tropicalis." Science (New York, NY) **328**(5978): 633-636.
Hiroi, Y., S. Kudoh, et al. (2001). "Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation." Nature genetics **28**(3): 276-280.
Hughes, M. K. and A. L. Hughes (1993). "Evolution of duplicate genes in a tetraploid animal, Xenopus laevis." Molecular biology and evolution **10**(6): 1360-1369.
Jamali, M., C. Karamboulas, et al. (2001). "BMP signaling regulates Nkx2-5 activity during cardiomyogenesis." FEBS Letters **509**(1): 126-130.
Kasahara, H., A. Usheva, et al. (2001). "Characterization of homo- and heterodimerization of cardiac Csx/Nkx2.5 homeoprotein." The Journal of biological chemistry **276**(7): 4570-4580.
Komuro, I. and S. Izumo (1993). "Csx: a murine homeobox-containing gene specifically expressed in the developing heart." Proc Natl Acad Sci USA **90**(17): 8145-8149.
Kontaraki, J. E., F. I. Parthenakis, et al. (2007). "Altered expression of early cardiac marker genes in circulating cells of patients with hypertrophic cardiomyopathy." Cardiovascular pathology : the official journal of the Society for Cardiovascular Pathology **16**(6): 329-335.
Krantz, I. D., R. Smith, et al. (1999). "Jagged1 mutations in patients ascertained with isolated congenital heart defects." American journal of medical genetics **84**(1): 56-60.
Kuo, C. T., E. E. Morrisey, et al. (1997). "GATA4 transcription factor is required for ventral morphogenesis and heart tube formation." Genes & Development **11**(8): 1048-1060.
Latinkić, B. V., S. Kotecha, et al. (2003). "Induction of cardiomyocytes by GATA4 in Xenopus ectodermal explants." Development (Cambridge, England) **130**(16): 3865-3876.
Lee, Y., R. Sultana, et al. (2002). "Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA)." Genome Res **12**(3): 493-502.
Li, L., C. J. Stoeckert, et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.
Lints, T. J., L. M. Parsons, et al. (1993). "Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants." Development **119**(2): 419-431.
Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nature genetics **21**(1 Suppl): 20-24.
Livak, K. (1997). "ABI Prism 7700 sequence detection system. User Bulletin 2." PE Applied Biosystems.
Lloyd-Jones, D., R. J. Adams, et al. (2010). "Heart disease and stroke statistics--2010 update: a report from the American Heart Association." Circulation **121**(7): e46-e215.

Luu-The, V., N. Paquet, et al. (2005). "Improved real-time RT-PCR method for high-throughput measurements using second derivative calculation and double correction." BioTechniques **38**(2): 287-293.
Lyons, I., L. M. Parsons, et al. (1995). "Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5." Genes & Development **9**(13): 1654-1666.
Maglott, D., J. Ostell, et al. (2007). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Research **35**(Database issue): D26-31.
McBurney, M. W., E. M. Jones-Villeneuve, et al. (1982). "Control of muscle and neuronal differentiation in a cultured embryonal carcinoma cell line." Nature **299**(5879): 165-167.
McClintick, J. N. and H. J. Edenberg (2006). "Effects of filtering by Present call on analysis of microarray experiments." BMC Bioinformatics **7**: 49.
McClintick, J. N., R. E. Jerome, et al. (2003). "Reproducibility of oligonucleotide arrays using small samples." BMC Genomics **4**(1): 4.
Meier, P., A. Finch, et al. (2000). "Apoptosis in development." Nature **407**(6805): 796-801.
Molkentin, J. D., C. Antos, et al. (2000). "Direct activation of a GATA6 cardiac enhancer by Nkx2.5: evidence for a reinforcing regulatory network of Nkx2.5 and GATA transcription factors in the developing heart." Developmental Biology **217**(2): 301-309.
Molkentin, J. D., Q. Lin, et al. (1997). "Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis." Genes & Development **11**(8): 1061-1072.
Monzen, K., I. Shiojima, et al. (1999). "Bone morphogenetic proteins induce cardiomyocyte differentiation through the mitogen-activated protein kinase kinase kinase TAK1 and cardiac transcription factors Csx/Nkx-2.5 and GATA-4." Molecular and Cellular Biology **19**(10): 7096-7105.
Nagao, K., Y. Taniyama, et al. (2008). "HIF-1alpha signaling upstream of NKX2.5 is required for cardiac development in Xenopus." The Journal of biological chemistry **283**(17): 11841-11849.
Oka, T., I. Komuro, et al. (1997). "Autoregulation of human cardiac homeobox gene CSX1: mediation by the enhancer element in the first intron." Heart and vessels **Suppl 12**: 10-14.
Oka, T., J. Xu, et al. (2007). "Re-employment of developmental transcription factors in adult heart disease." Seminars in cell & developmental biology **18**(1): 117-131.
Oliphant, T. E. (2006). Guide to NumPy. Manual.
Pandur, P., M. Läsche, et al. (2002). "Wnt-11 activation of a non-canonical Wnt signalling pathway is required for cardiogenesis." Nature **418**(6898): 636-641.
Patel, S., A. D. Leal, et al. (2005). "The homeobox gene Gax inhibits angiogenesis through inhibition of nuclear factor-kappaB-dependent endothelial cell gene expression." Cancer research **65**(4): 1414-1424.
Peterkin, T., A. Gibson, et al. (2005). "The roles of GATA-4, -5 and -6 in vertebrate heart development." Seminars in cell & developmental biology **16**(1): 83-94.

Peterkin, T., A. Gibson, et al. (2003). "GATA-6 maintains BMP-4 and Nkx2 expression during cardiomyocyte precursor maturation." The EMBO journal **22**(16): 4260-4273.
Petropoulos, H., P. J. Gianakopoulos, et al. (2004). "Disruption of Meox or Gli activity ablates skeletal myogenesis in P19 cells." Journal of Biological Chemistry **279**(23): 23874-23881.
Pikkarainen, S., H. Tokola, et al. (2004). "GATA transcription factors in the developing and adult heart." Cardiovascular Research **63**(2): 196-207.
Polack, S. S. (1949). "The xenopus pregnancy test." Canadian Medical Association journal **60**(2): 159-161.
Pontius, J. U., L. Wagner, et al. (2003). "UniGene: a unified view of the transcriptome." The NCBI Handbook. Bethesda (MD).
Ramakers, C., J. M. Ruijter, et al. (2003). "Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data." Neuroscience Letters **339**(1): 62-66.
Reecy, J. M., X. Li, et al. (1999). "Identification of upstream regulatory regions in the heart-expressed homeobox gene Nkx2-5." Development (Cambridge, England) **126**(4): 839-849.
Reiter, J. F., J. Alexander, et al. (1999). "Gata5 is required for the development of the heart and endoderm in zebrafish." Genes & Development **13**(22): 2983-2995.
Riazi, A. M., J. K. Takeuchi, et al. (2009). "NKX2-5 regulates the expression of beta-catenin and GATA4 in ventricular myocytes." PloS one **4**(5): e5698.
Rozen, S. and H. Skaletsky (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods in molecular biology (Clifton, NJ) **132**: 365-386.
Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science (New York, NY) **270**(5235): 467-470.
Schneeberger, C., P. Speiser, et al. (1995). "Quantitative detection of reverse transcriptase-PCR products by means of a novel and sensitive DNA stain." PCR methods and applications **4**(4): 234-238.
Schneider, V. A. and M. Mercola (2001). "Wnt antagonism initiates cardiogenesis in Xenopus laevis." Genes & Development **15**(3): 304-315.
Schott, J. J., D. W. Benson, et al. (1998). "Congenital heart disease caused by mutations in the transcription factor NKX2-5." Science (New York, NY) **281**(5373): 108-111.
Schwartz, R. J. and E. N. Olson (1999). "Building the heart piece by piece: modularity of cis-elements regulating Nkx2-5 transcription." Development (Cambridge, England) **126**(19): 4187-4192.
Sen, R. and D. Baltimore (1986). "Multiple nuclear factors interact with the immunoglobulin enhancer sequences." Cell **46**(5): 705-716.
Sepulveda, J. L., N. Belaguli, et al. (1998). "GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression." Molecular and Cellular Biology **18**(6): 3405-3415.
Sindelka, R., Z. Ferjentsik, et al. (2006). "Developmental expression profiles of Xenopus laevis reference genes." Developmental dynamics : an official publication of the American Association of Anatomists **235**(3): 754-758.

Sive, H. and R. Grainger (2000). Early development of Xenopus laevis: a laboratory manual, Cold Spring Harbor Laboratory Pr.

Skopicki, H. A., G. E. Lyons, et al. (1997). "Embryonic expression of the Gax homeodomain protein in cardiac, smooth, and skeletal muscle." Circulation Research **80**(4): 452-462.

Small, E. M. and P. A. Krieg (2003). "Transgenic analysis of the atrialnatriuretic factor (ANF) promoter: Nkx2-5 and GATA-4 binding sites are required for atrial specific expression of ANF." Developmental Biology **261**(1): 116-131.

Small, E. M., A. S. Warkman, et al. (2005). "Myocardin is sufficient and necessary for cardiac gene expression in Xenopus." Development (Cambridge, England) **132**(5): 987-997.

Sparrow, D. B., C. Cai, et al. (2000). "Regulation of the tinman homologues in Xenopus embryos." Developmental Biology **227**(1): 65-79.

Srivastava, D. and E. N. Olson (2000). "A genetic blueprint for cardiac development." Nature **407**(6801): 221-226.

Srivastava, D., T. Thomas, et al. (1997). "Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND." Nature genetics **16**(2): 154-160.

Stamataki, D., M. Kastrinaki, et al. (2001). "Homeodomain proteins Mox1 and Mox2 associate with Pax1 and Pax3 transcription factors." FEBS Letters **499**(3): 274-278.

Storey, J. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value." Annals of Statistics **31**(6): 2013-2035.

Tanaka, M., Z. Chen, et al. (1999). "The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development." Development (Cambridge, England) **126**(6): 1269-1280.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

Team, R. D. C. (2010). R: A Language and Environment for Statistical Computing. Manual.

Tonissen, K. F., T. A. Drysdale, et al. (1994). "XNkx-2.5, a Xenopus gene related to Nkx-2.5 and tinman: evidence for a conserved role in cardiac development." Developmental Biology **162**(1): 325-328.

Tu, C.-T., T.-C. Yang, et al. (2009). "Nkx2.7 and Nkx2.5 function redundantly and are required for cardiac morphogenesis of zebrafish embryos." PloS one **4**(1): e4249.

Ueyama, T., H. Kasahara, et al. (2003). "Myocardin expression is regulated by Nkx2.5, and its function is required for cardiomyogenesis." Molecular and Cellular Biology **23**(24): 9222-9232.

Vincentz, J. W., R. M. Barnes, et al. (2008). "An absence of Twist1 results in aberrant cardiac neural crest morphogenesis." Developmental Biology **320**(1): 131-139.

Wall, D. P., H. B. Fraser, et al. (2003). "Detecting putative orthologs." Bioinformatics **19**(13): 1710-1711.

Welch, B. L. (1947). "The Generalization of 'Student's' Problem when Several Different Population Variances are Involved." Biometrika **34**: 28-35.

Wheeler, D. L., T. Barrett, et al. (2008). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research **36**(Database issue): D13-21.

Wittwer, C. T., M. Gutekunst, et al. (1999). Method for quantification of an analyte, US Patent Office.

Wittwer, C. T., M. G. Herrmann, et al. (1997). "Continuous fluorescence monitoring of rapid cycle DNA amplification." BioTechniques **22**(1): 130-131, 134-138.

Xanthos, J. B., M. Kofron, et al. (2001). "Maternal VegT is the initiator of a molecular network specifying endoderm in Xenopus laevis." Development (Cambridge, England) **128**(2): 167-180.

Yin, Z., X. L. Xu, et al. (1997). "Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development." Development (Cambridge, England) **124**(24): 4971-4982.

Zaffran, S. and M. Frasch (2002). "Early signals in cardiac development." Circulation Research **91**(6): 457-469.

Zuscik, M. J., J. F. Baden, et al. (2004). "5-azacytidine alters TGF-beta and BMP signaling and induces maturation in articular chondrocytes." Journal of cellular biochemistry **92**(2): 316-331.

CURRICULUM VITAE

Marcus R. Breese

Education

2011 Doctor of Philosophy
Indiana University
Indianapolis, Indiana

Major: Biochemistry and Molecular Biology
Minor: Computer Science

2000 Bachelor of Science
Denison University
Granville, Ohio

Major: Biochemistry
Minor: Computer Science

Research experience

2009-2011 Visiting Scientist, Center for Bioinformatics and Computational Biology, Indiana University School of Medicine under Yunlong Liu, Ph.D.

Research areas: Established a data pipeline and tools for mapping and analysis of next-generation sequencing data, modeling of alternative splicing, and variation discovery in pooled genotyping samples.

2006-2011 Graduate student, Biochemistry and Molecular Biology, Indiana University School of Medicine under Howard J. Edenberg, Ph.D. (Continuation of prior work)

2001-2006 Graduate student, Biochemistry and Molecular Biology, Indiana University School of Medicine under Matthew W. Grow, Ph.D.

Research areas: Identification of targets of Nkx2-5 in *Xenopus laevis* using microarrays, cross-species annotation, laboratory data management, and spotted microarray fabrication.

1999-2000 Undergraduate research, Biochemistry, Denison University under Peter Kuhlman, Ph.D.

Research area: Analysis of co-variation in evolutionarily conserved proteins.

Publications

Breese, M.R., Grow, M.W., and Edenberg, H.J. (2011). CrossGene: Cross-species transcript homology and Gene Ontology annotation database. (in preparation)

Breese, M.R., Grow, M.W., and Edenberg, H.J. (2011). Identification of putative targets of Nkx2-5 in Xenopus laevis using gene expression analysis and cross-species annotation. *PLoS One*. (submitted)

Breese, M.R., Stephens, M.J., McClintick, J.N., Grow, M.W., Edenberg, H.J. (2003). Labrat LIMS: an extensible framework for developing laboratory information management, analysis, and bioinformatics solutions for microarrays. *Proceedings of the 2003 ACM Symposium on Applied Computing, Melbourne, FL.*

Presentations

2007 Indiana Bioinformatics Conference Indianapolis, IN
CrossGene: Transcript-centric cross-species gene homology database
Invited talk and poster

2005 Intelligent Systems in Molecular Biology Detroit, MI
BioNote: wiki-based knowledge base and collaborative environment
Poster, Lightning talk (BOSC)

2004 Indiana Bioinformatics Conference Indianapolis, IN
Labrat LIMS: Laboratory Information Management
Poster

2002 9th International Xenopus Conference Cambridge, UK
A New Microarray Designed for Genetic Pathway Identification in Xenopus Cardiovascular Development
Poster

1999 Argonne Undergraduate Research Symposium Argonne, IL
Algorithmic Analysis of Evolutionarily Conserved Combinations in Protein Sequences
Presentation

Honors

2011	Reviewer, BMC Genomics
2003-2008	Reviewer, ACM Symposium on Applied Computing
2005-2006	Graduate student representative to the faculty of Department of Biochemistry and Molecular Biology, Indiana University School of Medicine.
2004	Guest lecture, LIMS Bioinformatics class, Indiana University School of Informatics
1999	Anderson Summer Research Scholarship, Denison University

Selected Projects

NGSUtils
2010-2011

NGSUtils is a series of scripts and pipelines to aid in the processing and analysis of next-gen sequencing data (SOLiD and Illumina). Mapping was primarily done using the BFAST program. This package of programs includes pipelines for mapping reads to the genome and transcriptome, manipulation and filtering of BAM files, RPKM calculations, and various other scripts for managing and converting FASTA, FASTQ, SAM, and BAM files. Written in Python using the pysam library.

CrossGene
2006-2010

CrossGene is a web accessible database (http://crossgene.org) for finding gene orthologs between human, mouse, rat, chicken, frog, zebrafish, fly and nematode. It was done by constructing orthologous networks of reciprocal best BLAST matches between UniGene clusters for those organisms. Once the networks were assembled, GO term annotations from all members of a network were applied to all other members, allowing a poorly annotated organism like *Xenopus* use the information from better-annotated organisms. Web site and processing scripts written in Python.

BioNote
2005-2006

BioNote was a wiki designed for managing data within a lab environment. It served as an electronic lab notebook and general lab knowledge base for the Grow lab. Written in Java.

Labrat LIMS & Orderrat
2002-2005

Labrat LIMS was a project to managing and tracking data captured in processing microarray analysis. It was a web-accessible database that captured data in a flexible schema. It was also designed to capture data resulting from a defined laboratory workflow. Written in Java.
Orderrat was a companion webapp that managed order entry and tracking for the Grow and Edenberg labs for many years. Written in PHP.

Both Labrat LIMS and Orderrat were spun out to a startup for further development by Indiana University in 2004.

Skills

Research techniques

Laboratory	PCR, quantitative real-time PCR, RT-PCR, PCR primer design, molecular cloning, RNA/DNA extraction, RNA/DNA purification, agarose and poly-acrylamide gel electrophoresis, microscopy/imaging, *Xenopus* embryo culture, microinjection, microarray fabrication and analysis (spotted cDNA/oligonucleotide), high throughput screening, robotic sample preparation, laboratory information management (LIMS).
Bioinformatics	Next-gen sequencing: mapping, alignment, RNA-Seq, ChIP-Seq, CLIP-Seq, targeted resequencing, SNP identification, alternative splicing modeling. Experience with SOLiD and Illumina datasets. Microarray analysis: cDNA array design and analysis, Affymetrix GeneChip analysis. Gene identification, orthology, and ontology classification and prediction HPC cluster administration and programming (Torque/PBS)

Programming languages and computing environments

Languages	Python, Java, Bash, JavaScript, HTML, CSS, C++, C, PHP, Perl, R (in order of skill)
Libraries / tools	Numpy, Matplotlib, rpy (Python) Hibernate, Spring, Guice, Servlets, JSP, Ant (Java) MySQL
Operating systems	Mac OS X, Linux (various flavors), Microsoft Windows